UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Models of Social Networking and Information Diffusion in Future Internet Cyber-Physical Environments

Fabio Pezzoni

Anno 2014

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Models of Social Networking and Information Diffusion in Future Internet Cyber-Physical Environments

Autore:

Fabio Pezzoni _____

Relatori:

Prof. Enzo Mingozzi _____

Dott. Marco Conti _____

Ing. Andrea Passarella _____

Anno 2014
SSD ING-INF/05

Sommario

Negli ultimi anni, la grande diffusione di tecnologie ICT, es. dispositivi mobili e servizi di *social networking*, hanno portato alla proliferazione di reti elettroniche e comunità virtuali in cui i contenuti sono generati e propagati dagli utenti ICT. Poiché gli utenti interagiscono anche nel mondo fisico tramite relazioni sociali umane, le informazioni generate nel mondo virtuale possono produrre effetti nel mondo fisico e vice versa. Questo fenomeno, chiamato *cyber-physical convergence*, sta diventando un argomento di ricerca di rilievo per la progettazione e test di servizi avanzati di *social networking*.

In questa tesi vengono forniti nuovi fondamentali spunti riguardo la *cyber-physical convergence*, verificando che le proprietà delle reti sociali umane, formate nel mondo fisico, possono essere mappate direttamente sulle strutture sociali formate dagli utenti ICT nel mondo virtuale. Questo risultato ci permette di definire una nuova generazione di modelli di rete vengono catturate le caratteristiche chiave delle relazioni sociali umane sia del mondo virtuale che di quello fisico. Dato che le relazioni sociali sono una delle basi della comunicazione e degli schemi di scambio di informazioni tra utenti, questi modelli sono strumenti utili per progettare e valutare le performance di soluzioni tecniche *content-centric* per scenari in cui si verifica la *cyber-physical convergence*.

Un altro contributo di questa tesi, per la progettazione di servizi *content-centric*, è la definizione e la valutazione di nuove modelli di *information diffusion*. I nostri modelli sono in grado di riprodurre fedelmente i tipici schemi di diffusione delle OSNs, e che possono essere usati per studiare “in vitro” le performance dei servizi di *information diffusion* rispetto ai parametri chiave del sistema e del comportamento degli utenti.

Abstract

In the last years, the large diffusion of ICT technologies, e.g. personal devices and social networking services, led to the proliferation of electronic networks and virtual communities in which new content is generated and propagated by ICT users. Since users also interact in the physical world through human social relationships, the information generated in the cyber world can produce outcomes in the physical world and vice-versa. This phenomenon, called cyber-physical convergence, is becoming a prominent topic of research for the design of efficient Future Internet solutions. Indeed, such integration can be exploited for the design of efficient networking solutions for content dissemination and for the development and testing of advanced social networking services.

In this thesis we provide new fundamental insights about the cyber-physical convergence, verifying that the properties of the human social networks, formed in the physical world, can be directly mapped to the social structures formed by ICT users in the cyber world. This result allows us to define a new generation of network models where we capture the key characteristics of human social relationships both in the cyber and in the physical worlds. As social relationships are one of the basis of communication and information exchange patterns between users, these models are useful tools to design and evaluate the performance of content-centric technical solutions for cyber-physical converging scenarios.

Another contribution of this thesis, towards the design of content-centric services, is the definition and evaluation of novel information dissemination models. Our models are able to closely re-produce typical information diffusion patterns in OSNs, and can be used to understand “in vitro” the performance of information diffusion services with respect to key parameters of the system and users’ behaviour.

Questa tesi è dedicata alla mia famiglia

e agli amici di Malegno e della Valcamonica.

In qualsiasi parte del mondo mi trovi

vi porto sempre nel cuore.

Contents

List of Figures

List of Tables

1

Introduction

In the past decade, the advent of new Internet technologies introduced by the Web 2.0 and the proliferation of advanced personal mobile devices (e.g. smartphones, tablets and laptops) have drastically changed the way the information circulates. Nowadays, users of *information and communications technology* (ICT) are able to actively interact and collaborate with each other as creators of content, rather than being passive information consumers as in the past.

Interactions between ICT users can take place through direct or opportunistic communication between their personal mobile devices, leading to the formation of *electronic networks*. This kind of interactions requires the physical proximity of the devices whose range, even extended by opportunistic networking, is usually limited to small geographic areas. On the contrary, using the new Internet technologies, ICT users can smoothly interact within *virtual communities* that allow people to cross geographical and political boundaries in order to pursue mutual interests or goals. Various kinds of virtual communities have been developed (e.g. internet message boards, online chat rooms, virtual worlds), among which the *Online Social Networks* (OSNs) have recently become the most prominent (e.g. Facebook, Twitter and Google+).

The large diffusion of electronic networks and virtual communities led researchers to posit an integration process between the *physical world* of the individuals, and the *cyber world*, formed by the broad range of interactions between ICT users. Indeed, more and more often, content generated in the physical space produces outcomes in the cyber environment and, similarly, information generated in the cyber space has immediate influence on the physical environment [18]. This integration between the physical and the cyber worlds will become more and more

tight in the next future, thus playing a fundamental role in all the research areas of the Future Internet [17, 70].

The properties of *human social networks*, formed by social relationships between people in the physical world, fundamentally determine information diffusion in cyber-physical convergent environments. This is already clear in OSNs, where information flows following social links between users. For example, social relationships formed in the physical world can be translated into friendships in OSNs. Furthermore, an opinion expressed by an influential user in an OSN can change the public opinion and produce concrete effects in people's lives.

The aim of this thesis is to contribute to understand the interplay between social relationships and information diffusion in cyber-physical systems. This is central to design Future Internet services centred around content, which is one of the mainstream directions in the research community. Therefore, starting from the analysis of the properties of human social networks, we define network models that quantitatively capture the key features of social relationships. In fact, a better knowledge on the structure of the social networks can be exploited to design new efficient Internet technologies, for example regarding opportunistic networks for content dissemination. Social relationships can indeed be naturally translated into relationships between the users' devices predicting their interaction opportunities [19, 16]. Moreover, features of the social relationships can be used to characterise the communication channel, assigning different levels of privacy and reliability.

Finally, we aim at studying how information disseminates in complex, large-scale cyber-physical environments, and specifically at defining models of information diffusion defined, among other, by the parameters of social relationships between humans. We consider OSNs for validation purposes because social links in OSNs often present very similar properties with respect to "face-to-face" interactions between humans in the physical world, such as frequency of communication and level of trust. This convergence plays a fundamental role in the way information is disseminated in OSNs, also having a strong effect on the everyday life. Moreover, the importance of the social networking services in the generation and diffusion of new information, led researchers to investigate which other factors, in addition to the network structure, should be considered for designing optimised social networking services, for example socio-demographic factors, the popularity of the users and the online user behaviour [74, 84, 15, 31].

1.1 Contributions

In this thesis we verify the convergence between physical and cyber worlds through a structural analysis of the social networks in both environments. The results of this analysis are then used for the design of novel social network models that can be profitably used for the development and testing of Future Internet solutions. Finally, we contribute to the characterisation of information diffusion in OSNs with an extensive analysis of the user activity and the definition of a new information diffusion model.

1.1.1 Studying the Convergence between Human and Online Social Networks

The first contribution of this thesis is an extensive analysis on the convergence between the physical and the cyber worlds. In particular, we compare the structural properties of the human social networks (physical world) with those of the OSNs (cyber world) in order to find similarities. To characterise the human social networks we take advantage of the studies carried out in different disciplines, such as psychology and socio anthropology. On the contrary, as far as OSNs are concerned, we extract their structural properties processing a large data set from Facebook. In our analysis, we focus on the properties of the *ego networks* which are small portions of a social networks made up of an individual (called *ego*) along with all the social relationships he/she has with other people (called *alters*). The results demonstrate that ego networks in the two different worlds present very similar structural properties. Specifically, ego networks in Facebook share three of the most important features highlighted in physical environments: (i) they appear to be organised in four hierarchical circles; (ii) the sizes of the circles follow a scaling factor near to three; and (iii) the number of active social relationships is close to the well-known *Dunbar's number*. Assessing such similarity is very useful both for the design of Future Internet services and for the study of human social networks, since data from OSNs can be used in place of manually collected data. Moreover, this is one of the first studies on the characterising the ego networks in the cyber world.

1.1.2 Modelling the Human Social Networks

The strictly similarity found between human and online social network structures allows us to define a new generation of generative social network models that take advantage of the results obtained in the human social network domain. The use

of these results enables our models to characterise aspects of the human social behaviour that, to the best of our knowledge, are not considered in other solutions. In fact, we characterise the social links between pairs of nodes according to properties of the human social network. In particular, social links are organised into the observed hierarchy of social circles and their level of strength are obtained from a known distribution that is related to the inherent human social behaviour. Moreover, our models can reproduce other typical features of the social networks, that is the small world property and the presence of geographical constraints that makes physically close nodes more likely to have a social link. The models also present a parameter that can be tuned to modify the small world effect, producing networks with different features. The consistency of generated networks with the properties observed in both physical and cyber worlds makes our models suitable for the design of efficient Future Internet solutions. For instance, they can play a crucial role in simulative environment that require the use of realistic network structures in which the social links can be used to estimate the frequency of social interactions.

1.1.3 Analysing the Role of Human Sociality in Information Diffusion

The convergence between the physical and the cyber worlds led OSNs to gain an important role in the creation and diffusion of information between people. We contribute to the characterisation of the information diffusion in OSNs with an extensive analysis of a large data set of Twitter user activity. In the analysis we study the impact of different factors on the content propagation process. In particular, we pay special attention to the role of the content visibility, that needed more in-depth investigation. A key aspect of our analysis is the methodology we use for the measurement of the content visibility. Indeed, it is measured inferring the position of the content, at a given point in time, in the user message feeds. Using this measure, we observe that the position of the retweeted messages is distributed following a power-law function with coefficient 1.433. Analysing the behaviour of the most influential users, we observe that the visibility of the content and their popularity can not completely explain the influence they have in the network as latent factors emerge. Based on these results, we define an agent-based model of information diffusion that reproduces the behaviour of the users in Twitter, such that the impact of different parameters on information diffusion can be studied “in vitro”. Furthermore, proposed model can be profitably used for the development and testing of advanced social networking platforms.

1.2 Thesis Organization

The remainder of the thesis is organised as follows. In Chapter 2 we discuss the known properties of social networks in both physical and online environments, and besides, we survey works related to models of social networks and information diffusion. In Chapter 3 we process and analyse a Facebook data set, comparing the structural properties of OSNs with those of the human social networks. In Chapter 4 we propose two novel social network models that take advantage of the properties of the human social networks for the characterisation of the human social behaviour. In Chapter 5 we perform an extensive analysis of the information diffusion in Twitter. Finally, Chapter 6 draws our conclusion about models of social networks and information diffusion for the design of Future Internet solutions.

2

Background

Since before the widespread use of ICT, the characterisation of the *human social behaviour* has been a central topic in many different disciplines, such as psychology and social anthropology. Through *social relationships* each person constructs a sense of identify in relation to other people giving rise to *social interactions* that are the fundamental elements of any social phenomenon. There are several different conditions that support the establishment of a social relationship, for instance kinship, friendship, neighbourhood, membership in associations, etc. Once established, a social relationship can vary in intensity becoming stronger or weaker time after time and, possibly, it can cease to be.

Analysing the social relationships, researchers are able to model our society as a dynamic structure, called *social network*, in which individuals are interconnected by *social links* (i.e. social relationships). Through social networks, researchers can study the social groups in which people are organised and how they evolve over time. This kind of analysis is an extremely useful method for the comprehension of social phenomena in our society and, because of this, it has become a more and more important research activity, giving rise to a new specific discipline called *Social Network Analysis* (SNA).

Subsequently, the advent of OSNs fostered the analysis on social networks, since the abundance of online communication traces generated by social media allowed to overcome the problem of collecting large-scale social data sets that was posing strong limits to social sciences hitherto. Moreover, the high level of detail of the traces allowed to analyse concrete dynamic social phenomena, such as the diffusion of information.

2.1 The Strength of the Social Links

One of the most important aspect on the SNA is related to the definition of strength of a social link, usually called *tie strength*, and on the methods to measure it. A seminal study in this field has been carried out by the American sociologist Marc Granovetter which informally defined the tie strength as a linear combination of amount of time, emotional closeness (or intensity), intimacy and reciprocal services which characterize a social link [37]. He did not provide operational methods for estimating the tie strength, however he roughly classified social links into two categories: *strong ties* and *weak ties*, where the former represent important social relationships, such as close friends and relative, and the latter denote acquaintanceships or friends of a friend. According to his studies, weak ties are usually more in number than strong ties and, because of this abundance, they play a fundamental role in social networks [39]. For this reason, Granovetter assessed that researchers should not limit their model to deal with strong ties only but, for a fully comprehension of social phenomena, also weak ties must be considered.

In social anthropology, numerous measures of tie strength have been used or proposed. These include the emotional closeness [38, 57, 63], the frequency of contact [38, 57], the duration of the contact, the provision of emotional support and aid within the relationship [87], the mutual acknowledgement of contact [30], the social homogeneity [58] and the overlap of memberships in social groups [1]. These variety of measures indicates that tie strength is a concept that squares with different intuitions of the researchers.

The first extensive study on the measurement of the tie strength has been carried out by Peter Marsden and Karen Campbell in [60]. Using survey data on friendship ties, they applied multiple indicator techniques to construct and validate different measures. Among different indicators (emotional closeness, duration, frequency of contact, breadth of discussion topics and confiding) they concluded that the best measure of tie strength is given by the *emotional closeness* that refers to the feeling of people of being “close” to each other. Emotional closeness resulted to be free of contamination by other indicators and has been measured as a trichotomy: (i) acquaintance, (ii) good friend and (iii) very close friend. In case it is not possible to use the emotional closeness as measure of the tie strength, the frequency of contact can also be considered since there is a tight relation between the time invested in a relationship and its level of emotional closeness. However, in this case, values may require to be adjusted in order to avoid bias given by an overestimation of the tie strength between relatives, neighbours and co-workers.

A specific study on the effect of family ties on the estimation of the tie strength has been carried out by Roberts and Dunbar in [72]. They confirm that there may be differences in the time and cognitive resources required to maintain a social relationship at a certain level, depending on whether the existence of family ties. For example, maintaining a relationship at high level of emotional closeness requires a lot of time for both friends and relatives. On the contrary, for low levels of emotional closeness, family relationships require less invested time than the relationships with friends, thanks to sort of hidden family bonds.

Through the use of communication traces from Facebook, Arnaboldi et al. found consistency between the definition of tie strength given by Granovetter in [37] and a set of factors used to predict reference values of tie strength manually assigned by a sample of users to their social relationships [3]. Additionally, in [45], Jones et al. confirmed that the frequency of contact in online interactions is a good predictor of tie strength, using explicit tie strength evaluations given by a large set of participants.

2.2 The Ego Network Structure

Studies on the measurement of the tie strength allowed researchers to better characterise the properties of the social networks. In particular, numerous studies were conducted for analysing the social networks at the level of single persons. These studies are focused on the concept of *ego networks* that are simple social networks made up of an individual, called *ego*, along with all the persons, called *alters*, with whom the ego has a social link. Ego networks are a useful tool for studying the human social behaviour since they reveal how people organise and maintain their social relationships.

Even though it appears trivial, one of the most important properties of ego networks concerns the average maximum size. In fact, studies in anthropology and evolutionary psychology conducted by Robin Dunbar, demonstrated that the cognitive limits of the human brain constrain the number of social relationships an individual can actively maintain. Indeed, keeping a social relationship “active” requires a non negligible amount of cognitive and time resources, which are limited by nature. Studying the correlation between the neocortex size in primates and the dimension of their social group, he hypothesised that the average number of social ties an individual can actively maintain is approximately 150, widely known as *Dunbar's number* [24]. This results has been further confirmed by different studies on ego networks [92, 41].

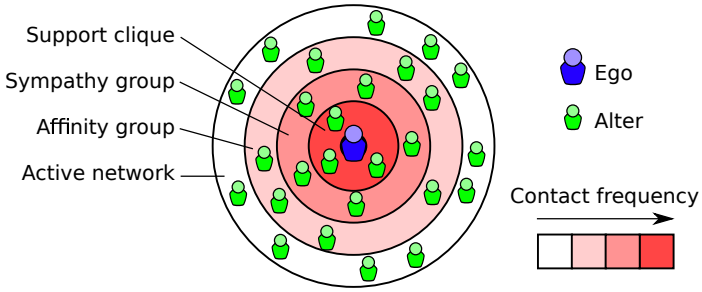

Figure 2.1: The ego network structure.

Considering the cognitive constrains and the different strength of the social links, Dunbar et al. observed the presence in ego networks of a hierarchical structure. This structure is composed by a series of concentric layers called *circles* in which alters are arranged on the base of the strength of their social links [24, 41]. In order to estimate the tie strength, researchers used the frequency of contact since it is easier to obtain than the emotional closeness, with which is positively correlated [72]. Studies revealed that the typical number of circles in an ego network is equal to four and that each of them can be characterised by the average size and the minimum frequency of contact of the social links [79]. Going from inner to outer circles, while the number of alters increases, the strength of the social links between the ego and the alters diminishes. This means that, typically, an ego has few very strong social relationships and a lot of weak ties as observed by Granovetter [37]. This structure is usually represented as an ego surrounded by as a series of circles, ordered by the strength of the social links, as depicted in Figure 2.1 [72].

The most inner circle of this structure is called *support clique* and contains the alters with whom the ego has the strongest social relationships, whom can be informally considered as *best friends*. They are the persons from whom ego seeks advice in case of severe emotional distress or financial disasters. The size of the support clique is, on average, limited to five members, usually contacted by the ego at least once a week. The second circle is called *sympathy group* and includes the alters whose social link with the ego is strong enough to be considered *close friends*. This circle has a size of around 15 members, contacted by the ego at least once a month. Most of the studies in ego networks focused on the two inner circles since their small size allows a practical collection and analysis of the data. For this reason, the support clique and the sympathy group have been better

characterised than external circles, for example classifying the alters in different categories as relatives, friends, neighbours, work colleagues, etc.

The third circle is the *affinity group* which contains alters connected by weaker ties than those in the sympathy group. Members of this circles corresponds to *casual friends* or extended family members who are, on average, limited to 50 persons [73]. Affinity group has been less characterised that other circles since the considerable number of members would require to collect information through interviews and surveys. For this reason a typical frequency of contact has not been defined yet. The last circle of an ego network is called *active network* and contains all the alters with whom the ego has an *active relationship*. By definition, an active relationship entails that the ego contacts the alter at least once a year. The active network includes all the other circles of the ego network and its size is limited by the Dunbar's number. Beyond the active network, some studies identified additional circles, called *mega-bands* and *large tribe* that include people with whom the ego has *inactive relationships*. This means that they are mere acquaintanceships and that the ego does not put any effort in them [79].

In [92], the authors analysed the structure of the ego network combining data from 61 studies on human grouping patterns. Using fractal analysis they confirmed, with high statistical confidence, the existence of a structure with 4 circles with preferred sizes of 4.6, 14.3, 42.6 and 132.5. This result is compatible with the approximate sizes described by Dunbar and, because of their accuracy, we will use these values as the reference sizes for the human ego network circles. In the same study, the authors highlighted that, also considering different data sets, the ratio between the size of two hierarchically adjacent circles is close to three. This ratio, called *scaling factor*, is considered to be an important property of the ego network structure and it has been hypothesised to be also related to the cognitive abilities of the human brain.

Apart from the socio anthropological studies on human ego networks, limited research work has been done to analyse the properties of online ego networks. In [36] the authors found a first evidence of the presence of the Dunbar's number in OSNs. They analysed a large-scale data set of Twitter communication data, finding that the average intensity of communication of each user towards all her friends, as a function of the number of social contacts of the user, shows an asymptotic behaviour, ascribable to the limits imposed by the Dunbar's number. Although this result gives a first insight on the constrained nature of online ego networks, there is still a lack of knowledge about all the other properties of the ego networks in online environments. Specifically, it is not clear if structures similar to those described in

socio anthropology literature about human ego networks could be found also in OSNs.

2.3 Social Network Macro-Level Properties

Seen from a macro-level perspective, social networks show some typical properties that have been observed in many different environments. Stanley Milgram, through his famous experiment, demonstrated the presence of the so called *small-world effect* in social networks [82]. According to this property, any two persons in the network, indirectly connected by chains of social links, have a short average distance. This is often identified as the *six degrees of separation* theory, for which everyone in a social network is six steps away. This fact directly influences the ability of the network to quickly spread information, ideas, innovations and so forth. According to the definition given by Watts and Strogatz [86], small-world networks presents two fundamental properties: (i) large clustering coefficient and (ii) a small average shortest path length.

Property (i) is strongly related with the triadic closure, that is a well-known concept in social network theory introduced by Granovetter in [37]. *Triadic closure* is defined as a property of the social networks for which, if a strong social tie exists between two pairs of nodes A-B and B-C, there is, with a high probability, a tie between the nodes C-A which closes the triangle. By definition, if a network complies with the triadic closures property, this guarantees a high level of clustering. According to Granovetter, the links in a network that do not take part in triangles are called *bridges*. They are mainly weak ties that have an important role in the social network structure as they connect socially distant parts of the network enabling to reach people and information not accessible via strong ties. As a consequence, the presence of bridges leads the average shortest path of a social network to be short, as required by property (ii).

The availability of OSNs communication data allowed to reveal the small world effect also in online environments. Specifically, it has been found in social graphs representing instant-message interactions between people [54, 22]. Moreover, in [61] the authors present a detailed analysis of the macro-level structural properties of a set of different OSNs, finding results in accordance with the properties of social networks observed in offline environments.

Another key aspect on the formation of human social network is the presence of *geographical constraints*. Indeed, for each person, it is more likely to have a social relationship with an individual who lives close to him, than to have a tie with a person who lives far away. This hypothesis is verified experimentally by Onnella

et al. in [66]. They analysed a huge data set of social interactions based on mobile phone calls in which each user is tagged with the geographical position where she probably lives. Plotting the frequencies of social ties between users which live at different distances, it emerges that the decay of the tie probability follows a power-law.

2.4 Information Diffusion in Social Networks

Based on the properties found in social networks, many different models have been proposed to characterise and replicate the dynamics of information diffusion processes. For a broad range of applications, two theoretical models of diffusion have been explored: the *linear threshold model* [39, 81] and the *independent cascade model* [33, 34]. These models assume similarly to what happens in a virus contagion. They produce a phenomenon called *cascading effect* in which information is propagated in the network following paths that are known in literature as *information cascades*. In the former model, given a node in the network, every infected neighbour contributes a certain weight and if the sum of the weights is greater than a threshold, the node is infected. The weights depend often on the edge strength between the node and its neighbours. In the latter model, each infected node is allowed one chance to infect a neighbour with some probability generally depending on the edge strength between the nodes.

Information diffusion models, like those cited, are widely used in marketing for studying the "word of mouth" effect in the promotion of new products [71, 34, 46, 27, 14, 2] and in economy for simulating the sudden and widespread adoption of various strategies in game-theoretic setting [9, 62, 59]. A useful application of these models is the *influence-maximisation*, that is an optimisation method for the selection of the set of seed nodes (i.e. nodes from which the diffusion process starts) which maximises the probability of diffusion in the network [23, 46, 14]. This approach is of particular interest for marketing, since these models could help reducing the costs of advertisement in social networks.

The advent of OSNs fostered the availability of large amount of information cascades data. The properties of the information cascades have been studied in different types on OSNs such as microblogging platforms like Twitter [31, 7, 91] and Facebook [77] and other specific Web 2.0 services, e.g. Flickr [13], blogs [55], Digg [80] and YouTube [78]. Most of the studies in literature propose models aimed at synthetically reproducing information cascades extracted from OSNs, like those presented in [35, 40, 53, 56]. These studies can lead to more effective and fair use

of these systems, suggest focused marketing strategies and provide insights into the underlying sociology.

In Chapter 5, we focus on the diffusion of information in Twitter since it is one of the most important networking platform nowadays (see Section 5.1.1 for a detailed description). In Twitter, information cascades can be obtained analysing the temporal sequence of retweets. The users who are able to originate, on average, large information cascades are considered the most influential, for the content they generate reaches a large number of users. Most of the studies in the literature about Twitter aim to discover which factors impact on the user influence and on the retweetability of the tweets (i.e. the probability for a given tweet to be retweeted). In [76], the authors examine a number of features that might affect retweetability of the tweets. They found that, amongst content features of the tweets, the presence of URLs and hashtags have strong relationships with the retweetability. Amongst contextual features, the number of followers and friends of the users, as well as the age of the accounts, seem to affect the user influence, while, the number of past tweets appears to be uncorrelated. This results have been confirmed in [8], in which authors found that the largest cascades in Twitter, tend to be generated by users who have a large number of followers and that the user influence appears to be rather constant in time. In [12] it is presented an in-depth comparison of three measures related to the user influence: the number of followers, retweets and mentions. This analysis revealed that there is a remarkable correlation between the number of followers and the number of retweet and mentions (i.e. the user influence), however popular users (i.e. who have high number of followers) are not necessarily influential.

In [31], the authors propose an information diffusion model to predict the information cascades in Twitter. The model relies on a set of latent variables that have to be trained using a training data set. Because of the computational complexity of the training process, the model is able to predict, with good accuracy, only the retweeting probabilities for the users that are one-hop away from the source. One of the main features of the model is the use of the *diffusion delay*, that is the time from the moment a friend of a user posts a tweet until the moment the user retweets it. The diffusion delay is fundamental in the analysis and modelling of information cascades in Twitter since, as reported in [50], half of retweeting occurs within an hour, and 75% under a day.

The trend of the diffusion delay is directly related with the phenomenon called *decaying visibility of the content*, that has been extensively studied by Hodes and Lerman in [42]. This phenomenon is based on the *principle of least effort* that links the effort required to perceive something to its *visibility*: high visibility contents take

little time and effort to be discovered, while low visibility contents require more time and energy to be perceived [48]. Since Twitter organises the tweets posted by user's friends in a chronologically ordered queue, the most recent tweets are at the top of the queue. Because users' attention is limited, they inspect only a finite portion of the queue, usually starting at the top [43]. Therefore, tweets residing at the top of the queue have the highest visibility, but visibility decays as new tweets arrive, pushing older ones farther down in the queue.

3

Structural Analysis of Online Social Networks

We are seeing a very significant process of integration between the physical world and the cyber world [18]. This is particularly evident in the area of social networks. In fact, Online Social Networks (OSNs) and human social networks definitely influence each other: people become friends in OSNs with individuals they also know "in the real life", while OSNs can be a means of reinforcing and maintaining social relationships existing in the physical world. Although several aspects are still under investigation, key properties regarding human social networks have been investigated quite extensively. On the other hand, the analysis of the properties of OSNs is much less advanced. The interplay between social interactions in the two types of networks is only partially understood and still under investigation [28, 10]. Moreover, the structural properties of OSNs, and their differences and similarities with human social networks are not yet fully understood.

In this chapter we focus on the latter aspect, providing a characterisation of structural properties of OSNs, based on the work in [4]. In particular we take advantage of a large data set that contains traces of communication between users in Facebook. We filter the data to obtain the frequency of contact of the relationships, then we check, by using different clustering techniques, whether structures similar to those found in human social networks can be observed. The results show a strikingly similarity between the social structures in human and online social networks. This similarity suggests that, even if the ways to communicate and to maintain social relationships are changing due to the diffusion of Online Social Networks, the way people organise their social relationships seems to remain unaltered. To further verify this result, we present a parallel study, based on the work in [5], in which we analyse a data set obtained from Twitter observing analogous properties.

3.1 Data Set Description

In this section we present a Facebook data set that contains a significant view of the online communication history of a high number of users. From this data set, we extract the frequency of contact for each social relationship that we use to characterise the ego networks of the users. Ego networks are analysed in this chapter to investigate the structural properties of the OSNs. Furthermore, the network graphs obtained from the data set are used in Chapter 4, to validate a social network model.

3.1.1 Platform Description

Facebook is the most used online social networking service in the world, with roughly 1.26 billion users as of 2013. Facebook was founded in 2004 and is open to everyone over 13 years old. Facebook provides several features to the users. Firstly, each user has a *profile* which reports her personal information and it is accessible by other users according to their permissions and the privacy settings of the user. Connected to her profile, the user has a special message board called *wall*, that reports all the asynchronous messages made by the user (*status updates*) or messages received from other users (*posts*). Posts (that include status updates) can contain multimedia information such as pictures, URLs and videos. Users can *comment* posts to create discussions around them. Comments have the same format as posts. To be able to access the personal page of other users a user must obtain their *friendship*. A friendship is a bi-directional relation between two users. Once a friendship is established the involved users can communicate with each other and view their personal information - depending on their privacy settings. The users can visualise the activity of their *friends* by using a special page called *news feed*.

Facebook provides also other mechanisms to communicate and maintain social relationships online, such as private messaging and voice and video calls. Another widely-used feature of Facebook is the *like* button, that allows people to express their favourable opinion about contents in Facebook (e.g. posts, pictures). Facebook is constantly adding new features to its service, providing always new way for the people to communicate and share content online.

3.1.2 Data Download

Although Facebook generates a huge amount of data regarding social communications between people, obtaining these data is not easy. In fact, publicly available

data have been strongly limited by the introduction of strict privacy policies and default settings for the users after 2009. Nevertheless, before that date most of the user profiles were public and another feature, that has been removed in 2009, allowed researchers to collect large-scale data sets containing social activity between users. In fact, before 2009 Facebook was built around the concept of *networks*. A network was a membership-based group of users with some properties in common (e.g. workmates, classmates or people living in the same geographic region). Each user profile was associated to a regional network based on her geographic location. By default, each user of a regional network allowed other users in the same network to access her personal information, as well as her status updates and the posts and the comments she received from her friends. Exploiting this characteristics of regional networks, some data sets has been downloaded, such as those described in [90]. The same authors made a pair of data sets crawled from Facebook regional networks on April 2008 publicly available for research[1]. The data set we used is referred as "Regional Network A" that has been studied in previous research work for different purposes [44].

The use of the regional networks feature allowed researchers to download large data sets from Facebook, however it entails some limitations that must be taken into account for our analysis. In fact, the considered data set contains information regarding the users within a regional network and the interactions between them only, excluding all the interactions and the social links that involve users external to this area. Therefore, assuming that for each user a part of his/her social relationships involve people who do not belong to the same network, this could lead to a reduction of the ego networks' size. Moreover, we do not have specific information about the completeness of the crawling process that should have downloaded only a sample of the original regional network. For example, in [89] the same crawling agent was used for downloading several other regional networks (not publicly available) collecting, on average, 56.3% of the nodes and 43.3% of the links. In absence of precise information, we assume that the sampling of the network affected nodes and links randomly.

3.1.3 Data Set Properties

The Facebook data set we use in this thesis consists of a *social graph* and four *interaction graphs*. These graphs are defined by lists of edges connecting pairs of anonymised Facebook user ids.

[1] `http://current.cs.ucsb.edu/facebook/`

Table 3.1: Statistics of the Facebook social and active graphs.

	social graph	**active graph**
# nodes	$3,097,165$	$1,171,208$
# edges	$23,667,394$	$4,357,660$
average degree	15.283	7.441
average clustering coefficient[a]	0.209	0.114
average shortest path	6.181	6.870

[a] Calculated as the average local clustering coefficient (Equation 6 in [64]).

The social graph describes the overall structure of the downloaded network. It consists of more than 3 million nodes (Facebook users) and more than 23 million edges (social links). An edge represents the mere existence of a Facebook friendship, regardless of the quality and the quantity of the interactions between the involved users. Basic statistics of the social graph are reported in Table 3.1.

The social graph can be used to study the global properties of the network, but alone it is not enough to make a detailed analysis of the structure of social ego networks in Facebook. Indeed, this analysis requires an estimation of the strength of the social relationships. To this aim, in Section 3.2, we leverage the data contained in the interaction graphs to extract the frequency of contact of the social links that can be used to estimate the tie strength.

Interaction graphs describe the structure of the network during specific temporal windows, providing also the number of interactions occurred for each social link. The four temporal windows in the data set, with reference to the time of the download, are: *last month*, *last six months*, *last year* and *all*. The latter temporal window ("all") refers to the whole period elapsed since the establishment of each social link, thus considering all the interactions occurred between the users. In an interaction graph, an edge connects two nodes only if an interaction between two users occurred at least once in the considered temporal window. The data set that we have used for the analysis contains interactions that are either Facebook Wall posts or photo comments.

In Facebook, an interaction can occur exclusively between two users who are friends. In other words, if a link between two nodes exists in an interaction graph, an edge between the same nodes should be present in the social graph. Actually, the data set contains a few interactions between users which are not connected in the social graph. These interactions probably refer to expired relationships or to

Table 3.2: Statistics of the Facebook interaction graphs.

	last mo.	**last 6 mo.**	**last year**	**all**
# nodes	$414,872$	$916,162$	$1,133,151$	$1,171,208$
# edges	$671,613$	$2,572,520$	$4,275,219$	$4,357,660$
average node degree	3.238	5.616	7.546	7.441
average edge weight	1.897	2.711	3.700	3.794

interactions made by accounts that are no longer active. To maintain consistency in the data set we exclude these interactions from the analysis. The amount of discarded links is, on average, 6.5% of the total number of links in the data set.

In Table 3.2 we report some statistics regarding the different interaction graphs. Each column of the table refers to an interaction graph related to a specific temporal window. The average degree of the nodes can be interpreted as the average number of social links per ego, which have at least one interaction in the considered temporal window. Similarly, the average edge weight represents the average number of interactions for each social link. Note that the measures reported in table can be influenced by the presence in the data set very low active users which are identified and discarded in Section 3.2.4.

The social graph contains some relationships with no interactions associated with them. These social links are considered as *inactive*. On the other hand we define as *active* all the relationships that have at least one interaction, that is to say the relationships included in the interaction graph "all". According to this classification, we define as *active graph* the sub-graph of the social graph obtained selecting the active social links and discharging the disconnected nodes.

A comparison between the properties of the social and the active graphs is reported in Table 3.1. As we can see in the table, the active graph is considerably smaller than the social graph. In fact, both many inactive nodes and edges where removed. Removed nodes represent either inactive users or users that have communicated with friends that do not belong to the examined regional network. Both social and active graphs present the typical properties exhibited by all the social networks studied in literature [64, 65]: high level of clustering coefficient and small average shortest path length (with respect to what one would expect on the basis of pure chance, given the observed degree distribution), thus the operation of removing inactive nodes and edges from the social graph do not affect the capability of the active graph to describe a typical real life social network.

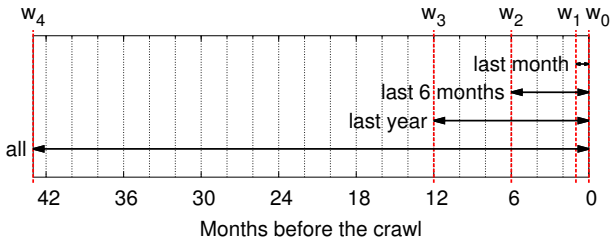

Figure 3.1: Temporal windows in Facebook data set.

3.2 Data Set Processing for Extracting Ego Networks

In order to characterise the social links in Facebook, we need to estimate the *link duration*, that is the time elapsed since the establishment of a social link. The link duration is needed to find the frequency of contact between the users involved in a social link, that is used to estimate the tie strength. In the literature, the duration of a social link is commonly estimated using the time elapsed since the first interaction between the involved users [32]. Unfortunately, the data set does not provide any indication regarding the time at which the interactions occurred. To overcome this limitation, we approximate the link duration leveraging the difference between the number of interactions made in the different temporal windows. Details on how we estimate the link duration and the frequency of contact between users in the Facebook data set are given in the next subsections.

3.2.1 Definitions

We define the temporal window "last month" as the interval of time (w_1, w_0), where $w_1 = 1$ month (before the crawl) and $w_0 = 0$ is the time of the crawl. Similarly we define the temporal windows "last six months", "last year" and "all" as the intervals (w_2, w_0), (w_3, w_0) and (w_4, w_0) respectively, where $w_2 = 6$ months, $w_3 = 12$ months and $w_4 = 43$ months. w_4 is the maximum possible duration of a social link in the data set, obtained by the difference between the time of the crawl (April 2008) and the time Facebook started (September 2004). The different temporal windows are depicted in Figure 3.1.

For a social relationship r, let $n_k(r)$ with $k \in \{1, 2, 3, 4\}$ be the number of interactions occurred in the temporal window (w_k, w_0). Since all the temporal windows in the data set are nested, $n_1 \leq n_2 \leq n_3 \leq n_4$. If no interactions occurred during a temporal window (w_k, w_0), then $n_k(r) = 0$. As a consequence of our definition

Table 3.3: Facebook classes of relationships.

class	time interval (in months)	condition
C_1	$(w_1 = 1, w_0 = 0)$	$n_1 = n_2 = n_3 = n_4$
C_2	$(w_2 = 6, w_1 = 1)$	$n_1 < n_2 = n_3 = n_4$
C_3	$(w_3 = 12, w_2 = 6)$	$n_1 \leq n_2 < n_3 = n_4$
C_4	$(w_4 = 43, w_3 = 12)$	$n_1 \leq n_2 \leq n_3 < n_4$

of active relationship, since $n_4(r)$ refers to the temporal window "all", $n_4(r) > 0$ only if r is an active relationship, otherwise, if r is inactive, $n_4(r) = 0$.

The first broad estimation we can do to discover the duration of social ties in the data set is to divide the relationships into different classes C_k, each of which indicates in which interval of time (w_k, w_{k-1}) the relationships contained in it started (i.e. the first interaction has occurred). We can perform this classification analysing for each relationship the number of interactions in the different temporal windows. If all the temporal windows contain the same number of interactions, the relationship must be born less than one month before the time of the crawl, that is to say in the time interval (w_1, w_0). These relationships belong to the class C_1. Similarly, considering the smallest temporal window (in terms of temporal size) that contains the total number of interactions (equal to n_4), we are able to identify social links with duration between one month and six months (class C_2), six months and one year (class C_3), and greater than one year (class C_4). The classes of social relationships are summarised in Table 3.3.

3.2.2 Estimation of the Duration of the Social Links

Although the classification given in Section 3.2.1 is extremely useful for our analysis, the uncertainty regarding the estimation of the exact moment of the establishment of social relationships is still too high to obtain significant results from the data set. For example, the duration of a social relationship $r_3 \in C_3$ can be either a few days more than six months or a few days less than one year. To overcome this limitation, for each relationship r in the classes $C_{k\in\{2,3,4\}}$ we estimate the time of the first interaction comparing the number of interactions n_k, made within the smallest temporal window in which the first interaction occurred (w_k, w_0), with the number of interactions (n_{k-1}), made in the previous temporal window in terms of temporal size (w_{k-1}, w_0). If $n_k(r)$ is much greater than $n_{k-1}(r)$, a large number of interactions occurred within the time interval (w_k, w_{k-1}). Assuming that these

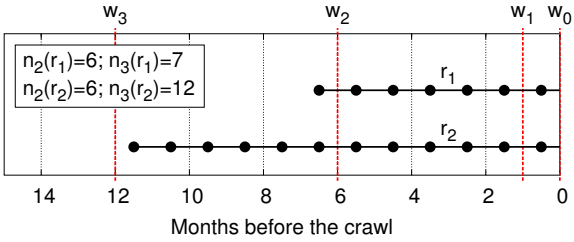

Figure 3.2: Graphical representation of two social relationships with different duration.

interactions are distributed in time with a frequency similar to that in the window (w_{k-1}, w_0), the first occurred interaction must be near the beginning of the considered time interval. On the other hand, a little difference between $n_k(r)$ and $n_{k-1}(r)$ indicates that only few interactions occurred in the considered time interval (w_k, w_{k-1}). Thus, assuming an almost constant frequency of interactions, the first contact between the involved users must be at the end of the time interval. The example in Figure 3.2 is graphical representation of this concept.

In the figure we consider two different social relationships: r_1, $r_2 \in C_3$. The difference between the respective values of n_2 and n_3 is small for r_1 and much larger for r_2. For this reason, the estimate of the time of the first interaction of r_1 must be near to w_2, while the time of the first interaction of r_2 results closer to w_3.

In order to represent the percentage change between the number of interactions n_k and n_{k-1}, we calculate, for each relationship $r \in C_k$, what we call *social interaction ratio* $h(r)$, defined as:

$$h(r) = \begin{cases} n_k(r)/n_{k-1}(r) - 1 & \text{if } r \in C_{k\in\{2,3,4\}} \\ 1 & \text{if } r \in C_1 \end{cases} . \tag{3.1}$$

If $r \in C_1$ we set $h(r) = 1$ in order to be able to perform the remaining part of the processing also for these relationships. The value assigned to $h(r)$ with $r \in C_1$ is arbitrary and can be substituted by any value other than zero without affecting the final result of the data processing. Considering that $n_k(r)$ is greater than $n_{k-1}(r)$ by definition with $r \in C_{k\in 2,3,4}$, the value of $h(r)$ is always in the interval $(0, \infty)$[2].

[2] In case $n_{k-1}(r) = 0$, we set $n_{k-1}(r) = 0.3$. This constant is the expected number of interactions when the number of interactions, within a temporal window, is lower than 1.

Employing the social interaction ratio $h(r)$, we define the function $\hat{d}(r)$ that, given a social relationship $r \in C_k$, estimates the point in time at which the first interaction of r occurred, within the time interval (w_k, w_{k-1}):

$$\hat{d}(r) = w_{k-1} + (w_k - w_{k-1}) \cdot \frac{h(r)}{h(r) + a_k} \qquad r \in C_k, \tag{3.2}$$

where a_k is a constant, different for each class of relationship C_k.

Note that the value of $\hat{d}(r)$ is always in the interval (w_{k-1}, w_k). The greater $h(r)$ - which denotes a lot of interactions in the time window (w_k, w_{k-1}) - the more $\hat{d}(r)$ is close to w_k. The smaller $h(r)$, the more $\hat{d}(r)$ is close to w_{k-1}. Moreover, the shape of the $\hat{d}(r)$ function and the value of a_k are chosen relying on the results about the Facebook growth rate, available in [89]. Specifically, the distribution of the estimated links duration, given by the function $\hat{d}(r)$, should be as much similar as possible to the distribution of the real links duration, which can be obtained analysing the growth trend of Facebook over time. For this reason, we set the constants a_k in order to force the average link duration of each class of relationships to the value that can be obtained by observing the Facebook growth rate. In the Appendix A we provide a detailed description of this step of our analysis.

3.2.3 Estimation of the Frequency of Contact

After the estimation of social links duration, we are able to calculate the frequency of contact $f(r)$ between the pair of individuals involved in each social relationship r:

$$f(r) = n_k(r)/\hat{d}(r) \qquad r \in C_k. \tag{3.3}$$

Previous research work demonstrated that the pairwise user interaction decays over time and it has its maximum right after link establishment [83]. Therefore, if we assessed the intimacy level of the social relationships with their contact frequencies, this would cause an overestimation of the intimacy of the youngest relationships. In order to overcome this problem, we multiply the contact frequencies of the relationships in the classes C_1 and C_2 by the scaling factors m_1 and m_2 respectively, which correct the bias introduced by the spike of frequency close to the establishment of the link. Assuming that the relationships established more than six months before the time of the crawl are stable, we set m_1 and m_2 comparing the average contact frequency of each of the classes C_1 and C_2, with that for the classes C_3 and C_4. Obtained values of the scaling factors are: $m_1 = 0.18$,

$m_2 = 0.82$. Setting $m_3 = 1$ and $m_4 = 1$, scaled frequencies of contact are defined as:

$$\hat{f}(r) = f(r) \cdot m_k \qquad r \in C_k. \tag{3.4}$$

3.2.4 Building and Selection of Ego Networks

In order to extract the ego networks from the Facebook data set, we group the relationships of each user into different sets R_e, where e identifies a specific ego. Since social links in the Facebook interaction graphs represent undirected edges, we duplicate each social link in the data set in order to consider it in both the ego networks of the users connected by it.

Since each ego in the data set has different Facebook usage, the calculated frequencies of contact are not directly comparable. For example, the same frequency of contact can represent, for different users, different levels of intimacy. To overcome this limitation, we normalise the frequencies of contact of each ego network in the interval between 0 and 1. This normalisation is essential to be able to compare the results of our analysis for different ego networks. Specifically, given an ego network R_e, we obtain the normalised frequency of contact $f_{norm}(r)$ of a relationship $r \in R_e$ by applying the following equation:

$$f_{norm}(r) = \frac{\hat{f}(r)}{\max\limits_{r' \in R_e} \hat{f}(r')} \qquad r \in R_e. \tag{3.5}$$

A high number of ego networks in the data set started just before the time of the crawl while other ego networks have a very low interaction level. The analysis could be highly biased by considering these outliers. Thus, we selected a subset of the available ego networks according to the following criteria. First of all we intuitively define as “relevant” the users who joined Facebook at least six months before the time of the crawl and who have made, on average, more than 10 interactions per month. We estimate the duration of the presence of a user in Facebook as the time since she made her first interaction. The new data set obtained from the selection of relevant ego networks contains $91,347$ egos and $4,619,221$ social links[3].

The average active ego network size after the cleanup is equal to 50.6. The discrepancy between this size and those found in other studies [41, 6, 36] is due to the fact that the data set contains a random sample of the social links (see

[3] $3,353,870$ bi-directional social links (without link duplication).

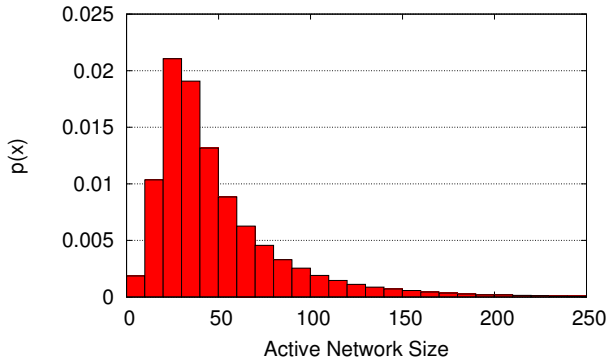

Figure 3.3: Active network size distribution.

Section 3.1.2). However, considering that in [89] the same crawling agent collected about 43% of the links, rescaled sizes appear to be compatible. Moreover, the active network size distribution (depicted in Figure 3.3) has a similar shape to those found in other analysis both in real and virtual environments [41, 6].

3.3 Aggregated Frequencies of Contact Analysis

As discussed in Section 2.2, the human social networks are characterised by the presence of concentric structures in each ego network. These structures are formed by nested layers, called circles, that include social links with different ranges of tie strength. In order to study the presence of the same social structures in OSNs, we analyse the frequencies of contact of the selected Facebook ego networks. The possible presence of social structures may be revealed by irregularities in the distribution of the frequency of contact since we can use it to quantify the tie strength [60, 72]. If the frequency of contact of an ego network appears uniformly distributed, this suggests the absence of any structure. On the contrary, if the frequency of contact appears clustered in different intervals, each of them may reveal the presence of a social layer.

The first attempt we make in order to check whether concentric structures are present in Facebook ego networks is to observe the complementary cumulative distribution function (henceforth CCDF) of the frequency of contact calculated aggregating all the frequencies of all the ego networks. We may expect this CCDF to have some kind of irregularities (e.g. jumps and plateaus) introduced by the possible presence of the clustered structure in the frequency of contact of the

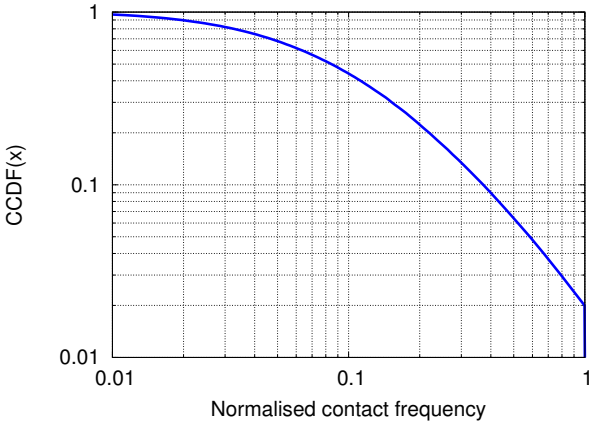

Figure 3.4: Aggregated CCDF of the normalised contact frequency for all the ego networks in the data set.

various ego networks. Yet, the CCDF (depicted in Figure 3.4) shows a smooth trend. This is not necessarily an indication of absence of clustered structures in individual ego networks, but it could be caused by the aggregation of the different distributions of the ego networks' frequency of contact. In fact, even if the single ego networks showed the circular hierarchical structure described in Section 2.2, the jumps between each circular cluster could appear at different positions from one ego network to another. This could mask jumps in the aggregated CCDF as we superpose the ego networks.

The CCDF of the aggregated frequency of contact shows a long tail, which can be ascribed to a power law shape. This may indicate a similarity between ego networks in human and online social networks, as studies in socio anthropology revealed that ego networks are characterised by a small set of links with very high frequencies of contact (corresponding to the links in the support clique). A power law shape in the aggregate CCDF is a necessary condition to have power law distributions in at least one ego network [68]. However, this is not a sufficient condition to have power law distributions in each single CCDF [67]. The presence of a long tail in the CCDF is not a conclusive proof of the existence of small numbers of very active social links in the individual ego networks. Therefore, to further investigate the online ego network structures we apply cluster analysis on each ego network looking for the emergence of layered structures.

3.4 Cluster Analysis Methodology

For each ego network, the frequencies of contact between ego and alters represent a set of values in a mono-dimensional space. Applying cluster analysis to mono-dimensional values does not require advanced clustering techniques, therefore we can consider standard widely-used methods such as k*-means clustering* and *density-based clustering*. Using k-means clustering, given a fixed number of clusters k, the data space is partitioned so that the sum of squared euclidean distance between the centre of each cluster (centroid) and the objects inside that cluster is minimised. In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set, that is usually considered to be noise [49].

3.4.1 k-means Clustering

In the first step of the cluster analysis we seek, for each ego network, the typical number of clusters in which the frequencies of contact could be naturally partitioned. In order to do this, we evaluate the goodness of different clustering configurations obtained applying the k-means clustering. This clustering method is defined as an optimisation problem that is known to be NP-hard. Because of this, the common approach for k-means clustering is to search only for approximate solutions. Fortunately, in the special case of mono-dimensional space, we can use an algorithm, called `Ckmeans.1d.dp`, able to always find the optimal solution efficiently [85]. Given a number of clusters k, the algorithm returns the optimal clustering configurations, those goodness is expressed in terms of *explained variance*, defined by the following formula:

$$VAR_{exp} = \frac{SS_{tot} - \sum_{j=1}^{k} SS_j}{SS_{tot}}, \tag{3.6}$$

where j is the j^{th} cluster, SS_j is the sum of squared distances within cluster j and SS_{tot} is the sum of squared distances of the all the values in the data space. Given a vector $\mathbf{X}$, the sum of squared distances $SS_{\mathbf{X}}$ is defined as $SS_{\mathbf{X}} = \sum_i (x_i - \mu_{\mathbf{X}})^2$, where $\mu_{\mathbf{X}}$ denotes the mean value of $\mathbf{X}$.

Given the number of clusters k, k-means clustering algorithms partition the space minimising the sum of squared distance within the clusters $\sum_{j=1}^{k} SS_j$. According to Equation 3.6, the optimal solution of the clustering, also provides the maximum value of the explained variance VAR_{exp}, since the sum of squared distances SS_{tot} is constant given the data space. In order to find the typical number

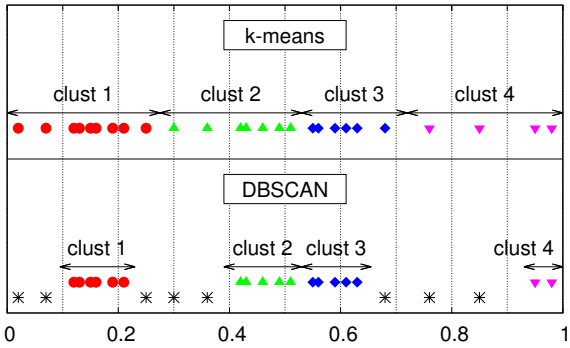

Figure 3.5: Example of different results obtained applying k-means and the iterative DBSCAN over a noisy data space, using $k = 4$.

of clusters k^*, we may calculate the optimal clustering for each k and then select the value that maximises VAR_{exp}. However, the value of VAR_{exp} increases monotonically with k, reaching its maximum when k is equal to the number of objects in the data space. Thus, there is a inherent overfitting problem. To avoid this overfitting we use an elbow method with a fixed threshold of 10% of the explained variance. Starting with $k = 1$, if, after adding a new cluster, the increment in terms of VAR_{exp} is less than 0.1, we take the current value of k as the typical number of clusters k^*. Otherwise, we iterate the procedure incrementing the value of k by one. This is a standard way to determine the typical number of clusters in a data set [47]. Hence, we apply this method to extract k^* and the cluster composition of all the ego networks the data set.

3.4.2 Density-Based Clustering

The results obtained with k-means could be potentially affected by the presence of noisy data. We use the notion of *noise* to define points in the data space with a very low density compared to the other points around them. Noise can affect the result of k-means in two different ways: (i) the presence of noisy points between two adjacent clusters could force the algorithm to discover a single cluster instead of two (the so called “single link effect” [49]); (ii) the presence of a large number of noisy points in the data set could lead k-means to detect clusters with a size larger than it should be according to a natural and intuitive definition of clustering (see Figure 3.5 for a graphical example). To verify that the noisy points in the data set do not excessively affect k-means we compare the results of the former with

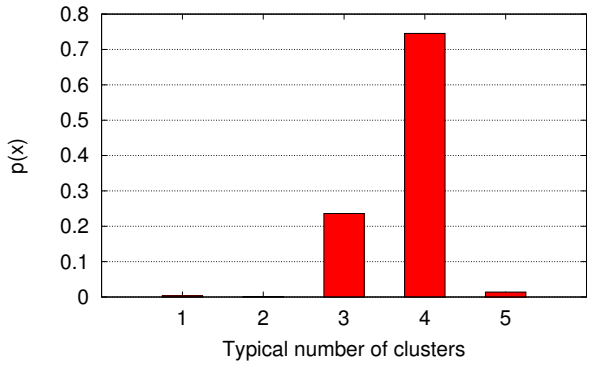

Figure 3.6: Distribution of k^* in Facebook ego networks.

the results of a density-based clustering algorithm called DBSCAN [29]. DBSCAN takes two parameters, namely ϵ and $MinPts$. If an object has more than $MinPts$ neighbours within an ϵ distance from it, it is considered a core object. A cluster is made up by a group of core objects (where two contiguous elements have a distance shorter than ϵ) and by the "border objects" of the cluster. Border objects are defined as non-core objects linked to a core object at a distance shorter than ϵ. For a more formal definition of density based clusters see [29]. Points with less than $MinPts$ neighbours within a distance equal to ϵ are considered noise by DBSCAN, and they are excluded from the clusters.

We iterate DBSCAN and we stop as we find a number of clusters equal to the number of clusters obtained by k-means. Hence, by comparing the results of k-means and DBSCAN in terms of cluster size we can verify that the former are valid and not influenced by noisy points. To allow noisy data to be identified by the iterative DBSCAN procedure we set the parameter $MinPts$ to be equal to 2. In this way isolated points are excluded from the clusters.

3.5 The Structure of Facebook Ego Networks

Using the iterative procedure based on the k-means algorithm (see Section 3.4.1), we obtain a distribution the typical number of clusters that ranges between 1 and 5, as shown in Figure 3.6. The average value of k^* is 3.76 ($\mathrm{SD} = 0.48$) and the median is 4. The presence of a typical number of clusters near to 4 in Facebook is the first indication of similarity between the findings in human ego networks and the ego networks in cyber environments. Since the amount of ego networks with

Table 3.4: # of ego networks and average active network size per each k^*.

k^*	# of nets	active net size	total contact freq.
1	315	1.50 [± 0.27]	15.21
2	107	3.81 [± 0.95]	18.83
3	$21,575$	34.42 [± 0.33]	26.96
4	$68,079$	55.23 [± 0.30]	35.64
5	$1,271$	77.74 [± 3.04]	37.87

k^* equal to 1 and 2 is negligible, compared with the total number of ego networks in the data set, we consider them as outliers and we exclude them in the subsequent part of the analysis.

Some properties of the ego networks for different values of k^* are reported in Table 3.4, in which the total contact frequency is expressed in terms of the total number of interactions per month made by ego. As can be seen in the table, the typical number of clusters and the active network size are positively correlated ($r = 0.25$, $p. < 0.01$). In this chapter, values in table between square brackets, indicate 95% confidence interval.

Considering the ego networks with $k^* = 4$, we apply the iterative DBSCAN procedure described in Section 3.4.2. The comparison between the inclusive circles found by k-means and DBSCAN on these ego networks and those found in human ego networks [92] are reported in Table 3.5. For each circle we show its average size and its ratio with the size of previous circle in the hierarchy (the scaling factor). We find that the results of k-means and DBSCAN are compatible in terms of circles size and their respective scaling factors. This means that k-means results are not highly influenced by noisy points (see Section 3.4). The discrepancy between the sizes of the support clique can be ascribed to the fact that DBSCAN considers isolated points as noise and, in many ego networks, the support clique could contain only one alter.

The scaling factors found by k-means in Facebook are strikingly similar to the findings in human ego networks (reported in Table 3.5 as "human"). Indeed, the average value of the scaling factors are equal to 3.12 in Facebook and 3.06 in human ego networks. In addition, we have rescaled the average size of the active network in Facebook to match that in human ego networks (132.50). The resulting ratio has a value of 2.63 that is compatible with the reported sampling of other networks obtained using the same crawling agent [89]. It is interesting to note

Table 3.5: Results for $k = 4$ of k-means (k-m) and DBSCAN (DB) on ego networks with $k^* = 4$.

	support clique	**sympathy group**	**affinity group**	**active network**
avg size (k-m)	1.84 [± 0.01]	6.36 [± 0.03]	18.68 [± 0.09]	55.48 [± 0.30]
scaled size (k-m)[a]	(4.70)	(15.31)	(44.77)	(132.50)
scaling factor (k-m)	-	3.45	2.94	2.97
min contact freq (k-m)	4.46	1.81	0.66	0.11
avg size (DB)	2.74 [± 0.01]	6.85 [± 0.04]	17.24 [± 0.10]	49.11 [± 0.40]
scaling factor (DB)	-	2.5	2.52	2.85
avg size (human)	4.6	14.3	42.6	132.5
scaling factor (human)	-	3.10	2.98	3.11

[a] Scaled size to match the active network in human ego networks.

that, scaling the size of other Facebook circles according to this ratio, they match very well the respective sizes in human ego networks.

In Table 3.5 the minimum contact frequency of the relationships within each circle is expressed in number of interactions per month. Using this variable, calculated averaging the results on all the ego networks, we are able to describe the circles of the discovered structure in terms of typical frequency of contact. Our results indicate that, in Facebook, the support clique contains people contacted at least $\sim$ *weekly*, the sympathy group $\sim$ *twice a month*, the affinity group $\sim$ *eight times a year* and the active network $\sim$ *yearly*. These results indicate that also the typical frequency of contact of the Dunbar's circles in Facebook appear to be very similar to that found in human ego networks.

As regards the ego networks with k^* equal to 3, it is interesting to notice that they do not have a counterpart in human ego networks. Their size is, on average, smaller than the size of ego networks with k^* equal to 4 and they show a lower rate of *Facebook usage*, defined by the total frequency of contact of each ego (see Table 3.4). We hypothesise that these ego networks have the same structure of the ego networks with k^* equal to 4, but the results of k-means could be influenced by the presence of too few social links. To prove this fact we apply k-means on these ego networks forcing $k = 4$ and we compare the results with those found on ego networks with $k^* = 4$. Table 3.6 reports the results of this analysis. The last two rows of the table ("% of human" and "% of $k^* = 4$") represent the percentage of

Table 3.6: Results for $k = 4$ of k-means (k-m) on ego networks with $k^* = 3$.

	support clique	**sympathy group**	**affinity group**	**active network**
avg size	1.62 [± 0.01]	4.14 [± 0.03]	11.9 [± 0.10]	34.63 [± 0.30]
scaling factor	-	2.56	2.87	2.91
min contact freq.	7.07	2.39	0.71	0.12
estimated size	(3.74)	(9.56)	(27.49)	(80)
% of human	81.30%	66.85%	64.53%	60.38%
% of $k^* = 4$	109%	65.08%	63.71%	62.42%

size of the obtained circles w.r.t. the size of the respective circles found in human ego networks and those found with k-means on the ego networks of the data set with $k^* = 4$.

Ego networks with $k^* = 3$ show a support clique with size near to the dimensions found in human ego networks (81.30%) and to that found by k-means on ego networks with $k^* = 4$ (86.18%). The dimensions of the other circles are noticeably lower. This result indicates that, in Facebook, users tend to have a set of core friends whom they contact frequently even if they have a lower rate of Facebook usage compared to the average. Nevertheless, the dimensions of the remaining circles are sensibly lower than the dimensions of the circles found in larger ego networks with higher Facebook usage. Still, the average scaling factor for the circles of the ego networks with $k^* = 3$ - equal to 2.78 - remains close to three, as an additional proof of the similarity between online and human ego networks.

The typical contact frequencies of the circles of ego networks with $k^* = 3$ are the following: the support clique contains people contacted at least $\sim$ *seven times a month*, the sympathy group $\sim$ *twice a month*, the affinity group $\sim$ *eight times a year* and the active network $\sim$ *yearly*.

As far as the ego networks with k^* equal to 5, we add them to the ego networks with k^* equal to 4 and we re-apply k-means on the resulting set, forcing $k = 4$. The results do not differ significantly (in terms of circle sizes and scaling factors) from the results found on ego networks with $k^* = 4$, reported in Table 3.5.

Table 3.7: Properties of ego network circles in Twitter.

	support clique	**sympathy group**	**affinity group**	**active network**
avg size	1.74 $[\pm 0.03]$	5.75 $[\pm 0.07]$	17.56 $[\pm 0.21]$	70.04 $[\pm 0.69]$
scaling factor	-	3.31	3.06	3.99
min contact freq.	17.28	6.00	1.77	0.20

3.6 Results Validation using a Twitter Data Set

To further validate the results obtained in the previous analysis we briefly present a parallel study in which we use the same techniques on a data set obtained from Twitter. Details of this study are presented in [5].

The data set we use for our analysis has been obtained in November 2012 using the same crawling agent we present in Section 5.1.2 that downloaded the data of $303,902$ Twitter users. Since we interested on the ego networks representative of human social behaviour, we implement an automatic procedure to exclude from the data set, all the accounts that use Twitter for different purposes that maintaining social relationships, for instance accounts representing companies, public figures, news broadcasters and bloggers. Moreover, we performed a refined selection of the ego networks to identify the most relevant set for our study. In particular, similarly to the selection performed in Section 3.2.4, we discarded too recent accounts (i.e. created less that 6 months before the time of the download) and the social relationships with a duration lower than one month. For the resulting $86,662$ we calculated the frequency of contact of the social links by dividing the number of replies sent by their duration of the friendship.

In order to highlight analogies and differences with the properties of the human ego networks structure and the results found analysing the Facebook data set, we use the same clustering technique described in Section 3.4.1. Applying $k-$means on the frequencies of contact we obtain, for each ego network, a hierarchical structure composed of a certain number of circles. Similarly to Facebook, also Twitter ego networks exhibit a typical number of circles close to what observed in physical environments. Indeed, the average number of circles is equal to 3.60 and its median is 4. Moreover, the size of the different circles and the scaling factors between them, presented in Table 3.7 are very close to what observed in Facebook and thus compatible to the properties of the human ego networks.

3.7 Discussion

We analysed a data set containing social interaction data collected from Facebook that have been processed to obtain the online ego networks of a large number of users. The results indicate that online ego networks show the same properties as those found in human social networks. In fact, we have found that the typical number of social circles of the online ego networks is equal to 4 and the scaling factor between hierarchically adjacent circles is very close to 3. These results are in line with the fundamental properties of human social networks found in physical environments. To the best of our knowledge this is the first indication of a convergence between the ego network structures of human and online social networks.

Looking in detail at the properties of the circles obtained from Facebook, we matched them with those defined in socio anthropology. The results indicate that the four circles we have found in Facebook are directly mappable with their physical equivalents. Moreover, in order to further verify this convergence, we presented a brief description of a similar analysis that relies on a Twitter data set. Since the results of this analysis are almost identical to what obtained using the Facebook data set, this can be considered a validation our conclusions.

Results of the presented analysis are useful to characterise the properties of human behaviour in cyber environment. The presence of similarities in the structure of human and online ego networks allowed us to estimate some properties of the former, that were unavailable in literature since they were impossible to measure with traditional sociometric techniques. These results can also be profitably used, from a technological point of view, to asses requirements for Future Internet communication services based on human sociality. In fact, relationships in online social environments could be automatically arranged in the observed hierarchical structure in order to simplify their maintenance over time. OSN services could leverage the differences between ego network circles to provide the users with different tools oriented to maintain relationships in specific circles. For example, privacy settings could be automatically adjusted according to the strength of the relationships in each circle.

4

Human Social Network Modelling

Human social behaviour is a key aspect for the development of Future Internet solutions, such as social networking environments. In particular, models of human social relationships are fundamental to characterise these systems and to study their performance. In this chapter we present two different generative network models for building synthetic human social networks where the known properties of the human social behaviour are accurately reproduced. In Section 4.1 we introduce a model, called *single-ego model*, that focuses on the generation of multiple independent ego networks [69, 20]. In Section 4.2 we extend the single-ego model defining a new one, called *multi-ego model*, that allows the generation of entire social networks in which ego networks are interconnected [21].

Proposed network models goes well beyond the binary approach, whereby edges between nodes, if existing, are all of the same type. Used algorithms set the properties of each social link, by incorporating fundamental results from sociology and social anthropology that we discussed in Chapter 2. Consequently, the synthetic networks they generate accurately reproduce the known ego network features (e.g. the size and the composition of the circles) and, in case of the multi-ego model, the macroscopic properties of the social networks (e.g. the diameter and the clustering coefficient). Thanks to the convergence between human and online social networks (discussed in Chapter 3), we compare generated networks with a large-scale social network data set, validating that the multi-ego model is able to produce graphs with the same structural properties of human social networks.

4.1 A Generative Model for Ego Networks

In this section we propose a generative network model, called *single-ego model*, that can be used to build synthetic ego networks following the properties summarised in Section 2.2. In contrast with the four-layer structure described in literature, the model considers just three circles: the support clique, the sympathy group and the active network. In fact, we consider the affinity group to be merged within the active network since no accurate information is currently available about its properties. In describing the model, we use sometimes the term *external part* of a circle in order to refer to the part of the circle not overlapped with its inner circles (e.g. the external part of the sympathy group is the part of the circle not overlapped with the support clique).

4.1.1 Overview

In the model the strength of the social links is measured in terms of *emotional closeness* since, as discussed in Section 2.1, it is the most predictive indicator of the tie strength. Exploiting the relation between emotional closeness and the time invested in a social relationship we map the maximum amount of cognitive resources an ego allocates for socialising with a maximum percentage of time it spends for this task. Therefore, each link is annotated with the amount of time the ego devotes to that social relationship, determining a constraint on the size of the ego network [41, 72].

In order to obtain the emotional closeness of each social link we start from a given distribution which can be derived from the empirical evidence collected in the social anthropology literature described in Section 2.2. As described in detail in Section 4.1.2, the distribution is partitioned according to the layered structure of ego networks, such that it is possible to identify the sectors of the distribution from where emotional closeness samples, related to the (external part of a) specific circle, must be drawn. We define a specific function that correlates emotional closeness and time spent on the relationship, as shown in Section 4.1.3. This function guarantees that we obtain, on average, ego networks with appropriate expected size.

Based on the empirical distributions available in the literature describing the sizes of the circles, we can also draw samples for the sizes of the support clique and sympathy group. Based on (i) the total time budget of the ego; (ii) the samples of the support clique and sympathy group sizes; (iii) the distribution of emotional closeness and the corresponding conversion function to time spent in a social relationship, we can generate ego networks. Given a tagged ego, we start by sampling

the sizes of the support clique and sympathy group assuming that they are linearly correlated, as suggested in [26]. Then, we sample the emotional closeness values for relationships in the support clique, then in the external part of the sympathy group, and finally in the external part of the active network. The emotional closeness associated to a social relationship reduces the time budget of the ego left for the rest of the social relationships in the network. The process stops when the time budget is zero.

The literature proposes different categorisations of relationships and alters: kin, friends, neighbours, work colleagues, etc. Our model only considers the kinship with the ego and the gender of the alters because there are many data available about these categories [41, 26]. Therefore, each relationship in the model is characterised by the type (kin or not-kin) and by the gender of the alter according to the composition of an average ego network.

4.1.2 The Algorithm

The pseudo-code of the algorithm used to generate ego networks is shown in Figure 4.1.

To generate an ego network, the algorithm exploits a set of functions (h_d, f_S, f_W, f_B, $f_{A,D}$ and f_E) and parameters (μ_l and m) whose characteristics are derived from the analysis of the human ego network properties, as discussed in Section 4.1.3. Specifically, most of these functions represent densities of random variables that characterise the properties of the ego networks, such as the size of the circles. They can be derived by fitting the data available in the anthropology literature presented in Section 2.2.

The first step is assigning a gender to the ego, hereafter denoted as g. The variable g takes the value M for male ego and F for female ego. It is sampled from a Bernoulli distribution $Ber(m)$ where m is the probability that $gen = \mathrm{M}$ (line 2–3). Next, we sample the sympathy group size s_{sym} from the known probability density function f_S (line 4), with average value μ_{sym}. We can then derive the size of the support clique, by exploiting the fact that it is linearly correlated with the size of the sympathy group. Specifically, we sample the ratio w between the sizes of the two circles from the density function f_W, and derive the corresponding value of the support clique size, as $s_{\mathrm{sup}} = s_{\mathrm{sym}} \cdot w$ (lines 5–6). We hereafter denote the expected value of s_{sup} as μ_{sup}[1].

[1] Since the probability density functions used in the model return continuous values, but circle sizes have to be natural numbers, values are rounded using the dithering method [75]. Moreover each negative value is converted into a zero.

1: **procedure** CREATESINGLEEGONET
2: $g \leftarrow$ EXTRACTFROM$(Ber\,(m))$
3: $ego \leftarrow$ CREATEEGO(g)
4: $s_{\mathrm{sym}} \leftarrow$ EXTRACTFROM(f_S)
5: $w \leftarrow$ EXTRACTFROM(f_W)
6: $s_{\mathrm{sup}} \leftarrow s_{\mathrm{sym}} \cdot w$
7: $bdg \leftarrow$ EXTRACTFROM(f_B)
8: $done =\leftarrow$ `False`, $tot \leftarrow 0$, $i \leftarrow 0$
9: **repeat**
10: $l \leftarrow$ SELECTCIRCLE$(i, s_{\mathrm{sup}}, s_{\mathrm{sym}})$
11: $a, d \leftarrow$ EXTRACTFROM$(f_{A,D|L=l,G=g})$
12: $e \leftarrow$ EXTRACTFROM$(f_{E|D=d}\ in\ (\mathtt{low}_{l,d}, \mathtt{up}_{l,d}))$
13: $t \leftarrow h_d\,(e)$
14: **if** $t/2 < bdg - tot$ **then**
15: $r \leftarrow$ CREATERELATIONSHIP(l, a, d, e, t)
16: ADDRELATIONSHIP(ego, r)
17: $tot \leftarrow tot + t$
18: $i \leftarrow i + 1$
19: **else**
20: $done \leftarrow$ `True`
21: **end if**
22: **until** $done$
23: **return** ego ▷ s_{net} is the final value of i
24: **end procedure**

Figure 4.1: Pseudo-code of the algorithm used by the single-ego model.

In the next step the algorithm assigns the time budget bdg. This amount is extracted from the probability density function f_B (line 7). Then, the algorithm has all the required values to generate all social links of the ego network, which is done in the main loop of lines 9-22. The total time currently spent on social relationships, tot, is updated after each relationship addition, together with the counter i, which represents the current size of the network (line 8). The algorithm generates social links starting from the support clique (the inner-most circle). Based on the values of the circle sizes s_{sup} and s_{sym}, the algorithm can determine the circle of the relationship it is generating (line 10). For each relationship, the algorithm determines the type d (K for kin or NK for non-kin) and the gender of the alter a (M or F as the variable g). The values are sampled from the joint probability mass functions $f_{A,D}$. Specifically, as discussed in Section 4.1.3, this density changes depending

on the circle l and on the gender of the ego, g. Therefore the values a and d are sampled from the density $f_{A,D|L=l,G=g}$ (line 11).

The value of emotional closeness associated with the social relationship is denoted with e. The density of e depends on whether we are generating a kin or non-kin relationship. These are denoted as $f_{E|D=\mathrm{K}}$ and $f_{E|D=\mathrm{NK}}$, respectively. Samples are drawn from these densities, by considering only the portion of the density corresponding to the specific circle we are populating, i.e. sampling from the interval $\left(\mathtt{low}_{l,d}, \mathtt{up}_{l,d}\right)$ related to the current circle l and the type of the current relationship d (line 12). The sample of emotional closeness is translated into the time spent on that relationship (t) through the function h_d. Again, we actually use two different functions, depending on the type (kin, non-kin) of the relationship, throughout referred to as h_{K} and h_{NK} (line 13).

In principle, the generated relationship should be kept in the ego network only if the total time spent by ego on social relationships considering the new one (i.e. $tot+t$) is less than or equal to the ego's budget bdg. Directly applying this condition (i.e. $t > bdg - tot$) would result in never achieving the target budget bdg, and ultimately, in the fact that the expected value of total time $\mathbf{E}[tot]$ would always be lower than the expected budget $\mathbf{E}[bdg]$, instead of being equal, as needed. To this end, the condition to accept the new relationship is relaxed, as follows: $t/2 < bdg - tot$ (lines 14–18).

The final value of the counter i at the end of the loop represents the size of the generated ego network, s_{net}. If the functions and the parameters of the model are defined satisfying the properties discussed in Section 4.1.3, the algorithm should generate, on average, ego networks with the expected size μ_{net}.

4.1.3 Parameters and Functions

In this section we define all the parameters and functions used by the model. For each of them, we justify the definition based on the results in the social anthropology literature summarised in Section 2.2.

Size of the Circles

In the literature there are different values for the circle sizes, often with significant differences. We use [92] as the main reference, as authors collected all the required data about circle sizes and extracted the mean value for each circle. Based on this work, we set the mean support clique size $\mu_{\mathrm{sup}} = 4.6$, the mean sympathy group size $\mu_{\mathrm{sym}} = 14.3$ and the mean active network size $\mu_{\mathrm{net}} = 132.5$.

Parameter m

Parameter m is the probability to have a male ego, that is $gen = \mathrm{M}$. We can reasonably assume that $m = 0.50$.

Function f_S

The sympathy group size density is analysed in [26]. Authors present a histogram based on real data collected in a measurement campaign, that can be fitted by a $Gamma$ distribution. As f_S must be consistent with the mean size of the sympathy group μ_{sym}, we obtained that f_S should follow a $Gamma(4.1, 3.49)$ distribution, which results in an average value of 14.3.

Function f_W

The ratio between the support clique and the sympathy group sizes is sampled using the density function f_W. Since we have set the mean sizes μ_{sup} and μ_{sym}, we define f_W as a Normal distribution with mean equal to $\mu_{\mathrm{sup}}/\mu_{\mathrm{sym}} = 0.3217$. We have no explicit information about the standard deviation of the distribution, however it can be experimentally approximated, using the scatter plot proposed in [26]. A good approximation is obtained by setting the standard deviation to half of the mean, therefore f_W is the density function of a Normal distribution $f_W = Normal(0.3217, 0.1608)$.

Function f_B

This function provides the density of the time spent by egos in social activities. There are no detailed studies on this distribution, but we know that its average value must be in the order of 20% of the yearly time of an individual [25]. Therefore we define f_B with a mean value equal to $8760 \cdot 0.2 = 1752$ where 8760 is the number of hours in a year. In this way the expected value of time budget is $\mathbf{E}[bdg] = 1752$.

As discussed in Section 2.2, the density function f_B directly influences the distribution of the network size. The distribution of the network size is analysed in [41], where a histogram based on collected real data is presented. By fitting this histogram, and translating it into the distribution of total time, we obtain that f_B should be the density of a Gamma distribution $f_B = Gamma(205.48, 8.5264)$.

Table 4.1: Composition of sympathy group.

	$g = \mathrm{M}$		$g = \mathrm{F}$	
$a = \mathrm{M}$, $d = \mathrm{K}$	2.28	(15.98%)	2.38	(16.64%)
$a = \mathrm{F}$, $d = \mathrm{K}$	2.47	(17.26%)	3.53	(24.72%)
$a = \mathrm{M}$, $d = \mathrm{NK}$	7.38	(51.61%)	2.02	(14.14%)
$a = \mathrm{F}$, $d = \mathrm{NK}$	2.17	(15.15%)	6.36	(44.51%)
sum	14.3	(100%)	14.3	(100%)

Functions $f_{A,D}$

The density functions $f_{A,D}$ determines the type (d) and gender (a) of the alter, given the circle l at which it is located in the ego network. A key reference for it is Dunbar and Spoors [26], where the composition of the sympathy group for male and female egos is investigated. Considering the average size μ_{sym}, that is independent of the gender of the ego, the resulting compositions are reported in Table 4.1. These values are used to define μ_{sym}. Moreover, in the same work, authors also study the support clique, observing that there are no significant differences between the compositions of the two circles in terms of type and gender of the alter. For this reason we can set $f_{A,D|L=\mathrm{sup}} = f_{A,D|L=\mathrm{sym}}$.

Finally, regarding the external part of the active network circle we can indirectly estimate its composition starting from results in [73]. Specifically, we set $f_{A,D|L=\mathrm{net}}$ with the results presented in Table 4.2.

Emotional Closeness Intervals and Functions f_E

Densities of emotional closeness have been experimentally characterised in [73]. Emotional closeness can be represented with a real value e in the interval $(0, 1]$. In principle, there should be a big component of the distribution at $e = 0$, corresponding to social relationship outside the active network. We do not consider this component, i.e. the density is conditioned to the fact that a social relationship exists. Fitting the empirical distributions presented in [72] for kin and non-kin relationships, we obtain the following densities: $f_{E|D=\mathrm{K}} = Gamma(0.2, 2.296)$ and $f_{E|D=\mathrm{NK}} = Normal(0.5, 0.172)$.

For the purpose of the algorithm presented in Section 4.1.2, it is also necessary to identify the boundaries on the density domains, corresponding to each circle of the ego network structure. We propose to identify the boundaries as follows. Let us neglect for a moment the distinction between kin and non-kin emo-

Table 4.2: Composition of active network circle (external part).

	$g = \mathrm{M}$		$g = \mathrm{F}$	
$a = \mathrm{M}$, $d = \mathrm{K}$	11.46	9.70%	17.35	14.68%
$a = \mathrm{F}$, $d = \mathrm{K}$	18.00	15.23%	17.18	14.53%
$a = \mathrm{M}$, $d = \mathrm{NK}$	52.50	44.41%	38.90	32.91%
$a = \mathrm{F}$, $d = \mathrm{NK}$	36.24	30.66%	44.78	37.88%
sum	118.2	100%	118.2	100%

tional closeness distributions, and let us assume that one distribution is sufficient to model emotional closeness. The average sizes of the circles naturally define the expected percentage of social links in each of the circles. For example, $\mu_{\mathrm{sup}}/\mu_{\mathrm{net}}$ defines the expected percentage of social links in the support clique. Assuming a random sampling from the emotional closeness distribution, the boundary identifying the support clique must be such that the fraction of samples greater than the boundary matches the expected percentage of social links in the support clique. In other words, the support clique boundary (throughout identified with $\mathtt{low}_{\mathrm{sup}}$) should be such that the Complementary Cumulative Distribution Function (CCDF) $1 - F_E(\mathtt{low}_{\mathrm{sup}})$ is equal to $\mu_{\mathrm{sup}}/\mu_{\mathrm{net}}$.

This line of reasoning can be generalised to the case where the emotional closeness distribution depends on the type (kin, non-kin) of relationship, d. The partitioning we obtain for kin relationships is represented in Figure 4.2 (note that the percentages indicated for circles in the figure are computed considering the social relationships in the external parts of the circles). Specifically, to identify the boundaries for each type of distribution, we need to know the mean proportion of kin for each circle. Using Equation 4.1 we obtain the probability k'_l to have a kin in the external part of a circle l (remember that m represents the fraction of male egos; we denote with $\mathcal{L}$ the set of the circles in a ego network).

$$k'_l = \sum_{a \in \{\mathrm{M},\mathrm{F}\}} \big(m \cdot f_{A,D|L=l,G=\mathrm{M}}\,(a, \mathrm{K}) + (1-m) \cdot f_{A,D|L=l,G=\mathrm{F}}\,(a, \mathrm{K}) \big) \qquad , \forall\, l \in \mathcal{L} \tag{4.1}$$

Using the values k'_l it is possible to obtain the probability of having a kin, k_l, in the whole circle l by the Equation 4.2, where c is a subcircle of l.

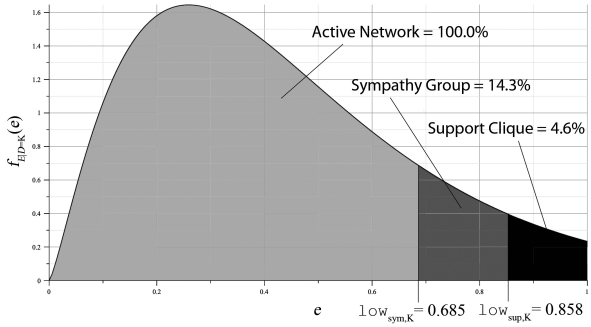

Figure 4.2: Distribution of the emotional closeness for kin.

$$k_l = \sum_{c \subseteq l} \frac{\mu'_c}{\mu_l} \cdot k'_c \qquad , \forall\, l \in \mathcal{L} \tag{4.2}$$

For example, the probability to have a kin in the whole network, k_{net}, is:

$$k_{\text{net}} = \frac{\mu'_{\text{net}} \cdot k'_{\text{net}} + \mu'_{\text{sym}} \cdot k'_{\text{sym}} + \mu'_{\text{sup}} \cdot k'_{\text{sup}}}{\mu_{\text{net}}} \tag{4.3}$$

Considering a type of relationship d, the probability to extract a value from $f_{E|D=d}$ in the interval $(\mathtt{low}_{l,d}, e_{max})$ related to a circle l, must be equal to the expected proportion of the network the circle l represents, considering only relationships with type d (e_{max} denotes the maximum possible value of emotional closeness, conventionally set to 1). Knowing the cumulative distribution functions F_E of the densities f_E, it is possible to calculate the boundaries, considering them as quantiles that satisfy the following equations:

$$F_{E|D=\text{K}}\left(\mathtt{low}_{\text{sup,K}}\right) = 1 - \frac{\mu_{\text{sup}} \cdot k_{\text{sup}}}{\mu_{\text{net}} \cdot k_{\text{net}}} \tag{4.4}$$

$$F_{E|D=\text{NK}}\left(\mathtt{low}_{\text{sup,NK}}\right) = 1 - \frac{\mu_{\text{sup}} \cdot (1 - k_{\text{sup}})}{\mu_{\text{net}} \cdot (1 - k_{\text{net}})} \tag{4.5}$$

$$F_{E|D=\text{K}}\left(\mathtt{low}_{\text{sym,K}}\right) = 1 - \frac{\mu_{\text{sym}} \cdot k_{\text{sym}}}{\mu_{\text{net}} \cdot k_{\text{net}}} \tag{4.6}$$

$$F_{E|D=\text{NK}}\left(\mathtt{low}_{\text{sym,NK}}\right) = 1 - \frac{\mu_{\text{sym}} \cdot (1 - k_{\text{sym}})}{\mu_{\text{net}} \cdot (1 - k_{\text{net}})} \tag{4.7}$$

For example, considering kin relationships and, again, the support clique circle, the limit $\mathtt{low}_{\text{sup,K}}$ defines an area in f_E whose size is equal to $\frac{\mu_{\text{sup}} \cdot k_{\text{sup}}}{\mu_{\text{net}} \cdot k_{\text{net}}}$ (the dark area in Figure 4.2) where $\mu_{\text{sup}} \cdot k_{\text{sup}}$ is the number of kin relationships in the support clique and $\mu_{\text{net}} \cdot k_{\text{net}}$ is the number of kin relationships in the whole network.

Considering the cumulative distributions F_E it is possible to solve the Equations 4.4−4.7, obtaining the limits of the intervals of emotional closeness: $\texttt{low}_{\text{sup,K}} = 0.8582$, $\texttt{low}_{\text{sup,NK}} = 0.8185$, $\texttt{low}_{\text{sym,K}} = 0.6852$ and $\texttt{low}_{\text{sym,NK}} = 0.7247$.

Functions h_d

The functions h_d (d denoting, again, the type of relationship kin, non-kin) map a value of emotional closeness to the time spent by the ego on that relationship. In general, we know from [41, 72] that those functions are monotonically increasing with the emotional closeness. We also know that $h_K(e) \leq h_{NK}(e)$, as kin relationships at a certain level of emotional closeness typically require less time than non-kin relationships, due to "embedded" familiar bonds. This difference actually fades out for high level of emotional closeness, such that it is reasonable to assume the following constraint:

$$h_{\text{K}}(e_{max}) = h_{\text{NK}}(e_{max}) \tag{4.8}$$

where e_{max} is the maximum level of emotional closeness.

Another constraint we have to impose is that the average total time spent on social relationships by an ego, according to the values provided by the functions h_d , equals the average time budget spent by the ego on social relationships, given by the density f_B, $\mathbf{E}[s_{bdg}]$. To have a reasonably simple expression for this constraint, we can compute the average time spent on a generic social relationship by the ego, according to functions h_d, and multiply it by the average size of the ego network, μ_{net} . This constraint can thus be expressed as:

$$\begin{aligned} \mu_{\text{net}} \cdot \int \big[& h_{\text{K}}(e) \cdot f_{E|D=\text{K}}(e) \cdot k_{\text{net}} \\ + & h_{\text{NK}}(e) \cdot f_{E|D=\text{NK}}(e) \cdot (1 - k_{\text{net}}) \big] \, \mathrm{d}e = \mathbf{E}[bdg] \end{aligned} \tag{4.9}$$

where the integral represents the average time spent by the ego on a generic social relationship (remember that k_{net} denotes the probability of kin relationships in the whole ego network).

The results presented in [72] and in [41] suggest that the h_d functions have an exponential shape. We thus define a general form for these functions, as $h(e) = c^e + t_0 - 1$. The parameter t_0 is the value $h(0^+)$, i.e. the minimum amount of time spent in a relationship in order to keep it active. The h_d functions for kin and non-kin relationships differ in the parameters c and t_0 . As previously noted, in general $h_K(e) \leq h_{NK}(e)$ must hold, therefore $t_{0,\text{K}}$ must be less than or equal to

Table 4.3: Circle sizes and time budget in synthetic ego networks.

	min	**max**	**avg**	**ref. value**
s_{sup}	0	45	4.62 $[\pm 0.01]$	4.6
s_{sym}	0	77	14.05 $[\pm 0.02]$	14.3
s_{net}	3	585	133.33 $[\pm 0.17]$	132.5
bdg	182.30	6074.64	1752.67 $[\pm 1.54]$	1752.0

$t_{0,\text{NK}}$. The literature does not provide any specific indications on these numbers. We thus set reasonable values for them, i.e. $t_{0,\text{K}} = 0.5$ and $t_{0,\text{NK}} = 2$. Finally, parameters c can be found by putting in a system Equations 4.8 and 4.9, and setting $\mu_{\text{net}} = 132.5$, $k_{\text{net}} = 0.2817$ and $\mathbf{E}[bdg] = 1752$. The system can be solved numerically, obtaining $c_{\text{K}} = 95.3275$ and $c_{\text{NK}} = 93.8275$. Therefore, the final form of the h_d functions is as follows:

$$h_{\text{K}}(e) = 95.3275^e - 0.5 \tag{4.10}$$

$$h_{\text{NK}}(e) = 93.8275^e + 1 \tag{4.11}$$

4.1.4 Results and Validation

To validate the algorithm for generating synthetic ego networks we have implemented it in a custom Java simulator and we performed $1.000.000$ run tests creating as many ego network graphs. Results are presented in Tables 4.3 and 4.4. Specifically, in Table 4.3 we present the ego network statistics for the circle sizes obtained in simulation. For each circle we show the minimum, maximum and average value, along with the 99% confidence interval. In the last column we also show the reference values for the average sizes of the circles, according to the anthropology literature. In Table 4.4 we show the composition of the ego network obtained in simulation, separately showing results depending on the gender of the ego (g) and alters (a_i), as well as the type of relationship between ego and alter (d_i). Again, average values are presented along with 99% confidence intervals and reference values.

The average network size converges to a value close to the expected value 132.5. Also the mean average of the sympathy group is very close to the reference value 14.3. In this case the small gap is due to the fact that time budget for social relationship may be over before the sympathy group is completed. This can happen for ego networks whose time budget sample is particularly low, or when the

Table 4.4: Composition of synthetic ego networks for male and female egos.

	$g = \mathrm{M}$ $(\mathbf{49.96}\%)$		$g = \mathrm{F}$ $(\mathbf{50.04}\%)$	
	avg	**ref. value**	**avg**	**ref. value**
$a_i = \mathrm{M}$, $d_i = \mathrm{K}$	13.72 $[\pm 0.03]$	13.74	20.03 $[\pm 0.03]$	19.73
$a_i = \mathrm{F}$, $d_i = \mathrm{K}$	20.46 $[\pm 0.04]$	20.47	20.99 $[\pm 0.04]$	20.71
$a_i = \mathrm{M}$, $d_i = \mathrm{NK}$	59.73 $[\pm 0.11]$	59.88	41.55 $[\pm 0.08]$	40.92
$a_i = \mathrm{F}$, $d_i = \mathrm{NK}$	38.35 $[\pm 0.07]$	38.41	51.82 $[\pm 0.09]$	51.14
sum	132.26 $[\pm 0.24]$	132.50	134.39 $[\pm 0.24]$	132.50

social relationships in the inner shells (support clique and sympathy group) are particularly strong (thus time demanding). In our tests, this happens in 3.17% of the runs.

The average size of the support clique matches perfectively its expected value. As in the case of the sympathy group, the time budget extracted can constrain the size of the circle. However, in case of the support clique, this happened only in 0.38% of the runs. We also checked the distributions of the circle sizes, comparing them with those found in the anthropology literature. For example, Figure 4.3 shows the distribution of the network size, which well matches empirical data. Moreover, Table 4.4 shows that the resulting composition of the ego networks is coherent in terms of type (kin, non-kin) and gender (male, female) with the empirical data. Male egos have smaller networks than females. This is due to female egos having a little more kin relationships which require less time that non-kin relationships.

4.2 A Generative Model for Entire Social Networks

In the previous section, we proposed a model for the generation of independent synthetic ego networks those satisfy well-known results in the field of social anthropology. In this section we present another model, called *multi-ego* model, that extends the previous one integrating several ego networks in a single synthetic human social network. Generated graphs has to satisfy both the properties of the single ego networks and, also, well-known macroscopic features such as the diameter and the clustering coefficient. The satisfaction of the ego network properties is guaranteed by the implementation of the single-ego model while, in order to reproduce the macroscopic features, the model relies the properties of social net-

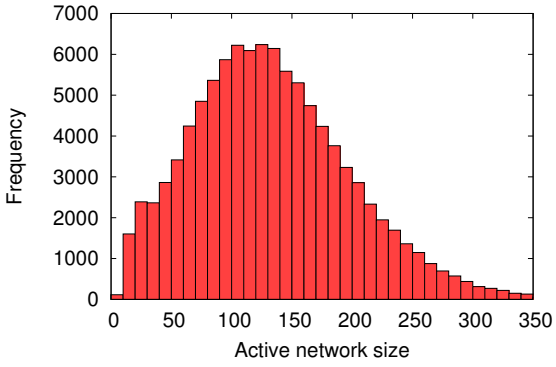

Figure 4.3: Synthetic ego network size distribution.

works discussed in Section 2.3, i.e. the triadic closure, the presence of bridges and geographical constraints [37, 66].

4.2.1 Overview

The model considers a human social network as a large group of individuals which are interconnected by social links. Intuitively, the procedure defined by the single-ego model can be applied to each of these individuals in order to generate its ego network. However, applying the single-ego procedure, we have to take into account that each new social link an individual adds to its ego network, also alters the ego network of the other individual involved in the relationship. This means checking, upon creation of a new link, that the properties of the involved ego networks are preserved. In detail, we have to check that (i) the size of the support clique, (ii) the size of the sympathy group, and (iii) the total budget of time remain consistent. Moreover, in order to generate complete ego networks we have to take into account the additional properties described in Section 2.3, i.e. triadic closure, presence of bridges and geographical constraints.

A new social link can be established either exploiting the triadic closure property or creating a bridge. The strategy to be used is randomly selected based on a given probability. In case the triadic closure strategy is selected, the procedure tries to close a triangle, that is, given an origin node, it selects a node at a distance of 2 hops as link's destination, favouring strong tie hops. On the contrary, in case the procedure follows the bridge creation strategy, the destination node is chosen randomly. In both cases geographical constraints have to be respected. In order

```
 1: procedure CREATESOCIALNETWORK(n, p, f_D, f_B, f_S, f_W, f_E, h)
 2:     for i ← 1, n do
 3:         i ← CREATEEGO(f_B, f_S, f_W)
 4:         i.pos ← EXTRACTFROM(Uniform(−1, 1))
 5:         V ← V + i
 6:     end for
 7:     for all circle l ∈ {sup, sym, net} do          ▷ maintaining the ordering
 8:         while OPEN(V, l) is not empty do
 9:             i ← random select in OPEN(V, l)
10:             if RAND() < p then
11:                 j ← CLOSURESELECT(i, f_D, OPEN (V, l))
12:             else
13:                 j ← BRIDGESELECT(i, f_D, OPEN (V, l))
14:             end if
15:             r ← NEWSOCIALLINK(i, j)
16:             r.e ← EXTRACTFROM(f_E in (low_l, up_l))
17:             update E, i.size, j.size, i.dbg and j.bdg
18:         end while
19:     end for
20:     return V, E
21: end procedure
```

Figure 4.4: Pseudo-code of the algorithm used by the multi-ego model.

to do this, we incorporate geographical information into the nodes, associating to them random locations in a virtual space. Whatever strategy to create links is selected, the model guarantees that the probability to have a social link between two nodes is proportional to a power law of the distance between them. Remember this is consistent with empirical results in the literature [66].

4.2.2 The Algorithm

The pseudo-code of the algorithm used for generating synthetic human social network graphs is shown in Figure 4.4. The input required by the algorithm consists of: (i) the number of nodes in the network n; (ii) the probability p to create a new social link using the triadic closure property rather than creating a bridge; (iii) the power-law distribution function f_D which gives the probability to establish a social link between nodes at a specific distance; (iv) the parameters used to define the structure of the single ego networks f_B, f_S, f_W, f_E, h, as required by the single-ego model (see Section 4.1.3).

22: **procedure** BRIDGESELECT($i, f_D, Open$)
23: $J \leftarrow Open - Neighbours(i) - i$
24: **if** J is not $empty$ **then**
25: $j \leftarrow$ select in J with $P \propto f_D(dist(i,j))$
26: **return** j
27: **else**
28: **return** failure (close node i)
29: **end if**
30: **end procedure**

Figure 4.5: Pseudo-code of the bridging procedure.

In the first part of the algorithm we create and initialise each node i in the network as an ego (lines 2-6). For each node we first call the procedure CREATEEGO which sets the size of the sympathy group $i.s_{\mathrm{sym}}$ and the size of the support clique $i.s_{\mathrm{sup}}$. It also assigns the budget of time $i.bdg$ and initialises the counter $i.size$ which is then used to keep track of the total size of the ego network (line 3). We also assign a geographical position of the ego ($i.pos$) which is randomly selected in a given space which, without loss of generality, we assume mono-dimensional, circular and included in the interval between -1 and 1. This definition guarantees that the distance between any pair of nodes is between 0 and 1 (line 4). Finally, each generated ego is included in the set V (line 5).

After the initialisation of the egos, we start adding social links to the network. First, we create all the social links belonging to all the support cliques, then we continue with the sympathy groups (external part), and finally we add the links of the active networks (external part) (line 7-17). Given the circle l we are populating, the creation of a new social link between two nodes i and j starts with the selection of the node i, drawn randomly from the nodes labelled as *open* (line 9). An "open" node is an ego whose population of the current circle l is not yet completed[2]. The selection of the nodes involved in a new social link from the open node set OPEN (V, l) guarantees the preservation of the ego network properties. The fundamental part of the algorithm is the selection of node j. We use two different strategies: (i) the *triadic closure* mechanism (procedure CLOSURESELECT) and (ii) the *bridging* (procedure BRIDGESELECT). The former strategy is chosen with a probability given by the parameter p, while the latter with probability $1 - p$ (lines 10-14).

[2] In case the current circle l is the support clique or the sympathy group, an ego i is open if its ego network size $i.size$ has not reached the thresholds $i.s_{\mathrm{sup}}$ or $i.s_{\mathrm{sym}}$ respectively. In case l is the active network, i is open if it has not exhausted its time budget $i.bdg$.

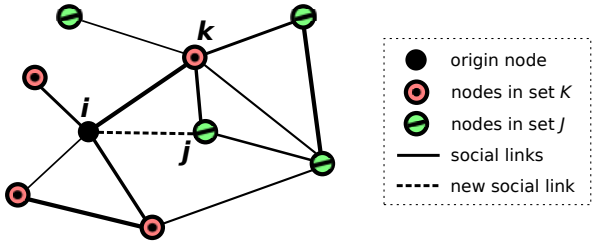

Figure 4.6: Triadic closure strategy.

In Figure 4.5, we show the pseudo-code of the *bridging*, that is the simple strategy. It consists in the selection of a node j in order to create a new bridge with the current node i. We extract the node j from the open egos in the network for the current circle l, excluding the nodes already connected to i, taking into account the geographical constraints. The probability to select a node j is thus proportional to the value of the power-law function f_D (discussed in Section 4.2.3), given the distance $dist(i, j)$ between i and j. Formally,

$$P(j) \propto f_D(dist(i,j)) \qquad j \in \textsc{Open}\,(V, l) - Nei(i) - i \tag{4.12}$$

where $Nei(i)$ is the set of one-hop neighbours of node i.

If each node in the network, not connected to node i, is *closed* (not open), node j can not be selected. In this case node i is forced to be closed. We have experimentally checked that this circumstance occurs just in a negligible number of cases and that the overall results are not affected.

In using the *triadic closure* strategy, represented in Figure 4.6 and whose pseudo-code is shown in Figure 4.7, we select a node j at a distance of 2 hops from the current node i in order to close a triangle. For the selection of the node j, according to the definition of triadic closure, we follow with a higher probability the strong ties. We first select the set K of the neighbours of i. From this set, we extract an intermediate node k with a probability that is proportional to the tie strength e_{ik} between i and k multiplied, in order to satisfy the geographical constraints, by a function of the distance $dist(i, k)$ (Equation 4.13). Given the intermediate node k and the current circle l, we define the set J as the set of open neighbours of k, with respect to l, excluded node i and its neighbours. From the set J we extract node j using the same method used for the selection of node k, considering the social relationship between k and j (Equation 4.14).

31: **procedure** CLOSURESELECT($i, f_D, Open$)
32: $K \leftarrow Neighbours(i) \cap Open$
33: **while** K is not $empty$ **do**
34: $k \leftarrow$ select in K with $P \propto e_{ik} \cdot \sqrt{f_D(dist(i,k))}$
35: $J \leftarrow (Neigbours(k) - i) \cap Open$
36: **if** J is not $empty$ **then**
37: $j \leftarrow$ select in J with $P \propto e_{kj} \cdot \sqrt{f_D(dist(k,j))}$
38: **return** j
39: **else**
40: $K \leftarrow K - k$
41: **end if**
42: **end while**
43: **return** failure (use bridging)
44: **end procedure**

Figure 4.7: Pseudo-code of the triadic closure procedure.

$$P(k) \propto e_{ik} \cdot \sqrt{f_D(dist(i,k))} \qquad k \in K = Nei(i) \tag{4.13}$$

$$P(j) \propto e_{kj} \cdot \sqrt{f_D(dist(k,j))} \qquad j \in J = Nei(k) \cap \text{OPEN}\,(V,l) - i - K \tag{4.14}$$

If the set J is empty we go a step backward and we select a different node k. If, for each k chosen, it is not possible to define a non-empty set J, the procedure fails and the algorithm recovers selecting j using the bridging. Bridging is also used in case node i has not neighbours, i.e. the set K is empty.

The function of the distance we use in Equations 4.13 and 4.14 is defined as the square root of the function f_D. This definition guarantees that the geographical distance between connected nodes in the final network follows the power-law rule defined in f_D. In Figure 4.8 we show a comparison between a given function f_D and the geographical distances obtained using this algorithm.

After the selection of node j, a new social link r between nodes i and j is created (line 15). Its emotional closeness $r.e$ is extracted from the density function f_E in the same manner as in the single-ego model (line 16). Then, we update the network adding the new social relationship r to the set of links E. We also update the egos i and j, in terms of the ego network sizes ($i.size$ and $j.size$ respectively) and of the residual budget of time ($i.dbg$ and $j.bdg$ respectively) (line 17). It is worth noting that this update can determine the transition of a node from the open to the closed state, with respect to the current circle l.

For each circle l, we generate and add new social links until there are open nodes available. When the set of the open nodes is empty, the procedure switches to the next circle until all the three circles are completed.

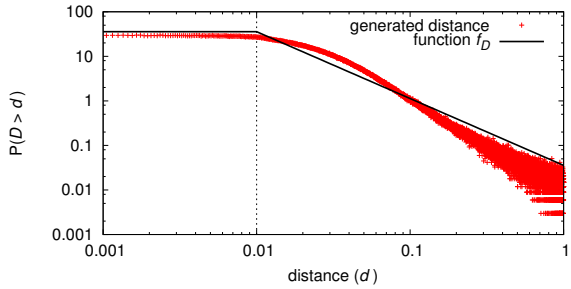

Figure 4.8: PDF of the generated distance and the function f_D ($\alpha = 1.5$, $d_{min} = 0.01$).

4.2.3 Geographical Distance Distribution Function

According to the results presented in [66] and summarised in Section 2.3 the probability of contact between two users at a certain distance follows a power-law of the form $P(d) \propto d^{-\alpha}$. In order to obtain a related probability density function f_D we have to introduce a threshold d_{min} from which the power-law hods. Moreover it has to be defined for the range of values of d, which is the interval $(0, 1)$. The function, shown in Figure 4.8, is thus defined as:

$$f_D(d) \propto \begin{cases} d_{min}^{-\alpha} & \text{for } 0 < d < d_{min} \\ d^{-\alpha} & \text{for } d_{min} < d < 1 \end{cases} \tag{4.15}$$

Experimental results in [66] suggest that $\alpha = 1.5$. On the contrary, a value for d_{min} cannot be set in general since it strongly depends on the geographical space we consider and on the geographical distribution of the sampled population. Note that, given the number n of nodes in the network, since they are equally distributed in the space, $n \cdot d_{min}$ is the average number of nodes within the distance d_{min} from any given position. Thus, given a node in the network, the closest $n \cdot d_{min}$ nodes (on average) have the same highest-probability to be selected as destination of a social link. This parameter impacts on the clustering coefficient of the network, as we highlight in Section 4.2.5.

4.2.4 Reference Network Properties

The reference network we use for the validation of our model is obtained from the Facebook data set we described in Section 3.1. As discussed in Section 3.2.4,

we dismissed some users from the original data set since they were not considered relevant, either for having too few interactions, or because they had joined Facebook just before the beginning of the data collection period. The new data set obtained from the selection of relevant egos and the social links between them contains $91,347$ users and $1,264,658$ social links which are labelled with the normalised frequency of contact between users.

Relevant properties of the reference network are reported in the second column of Table 4.5. The high clustering coefficient (with respect to random networks) and the short average path length prove that the reference network is "small-world". Analysing the properties summarised in the table we have to take into account that, for technical reasons (e.g. the discard of not relevant nodes), the data set captures just a random sub-sample of the social links on the crawled Facebook networks and some of the indexes are influenced by the sampling, i.e. the average degree and the average path length. If we had the complete network, we would most likely find a higher average degree and a shorter path length. On the contrary, the clustering coefficient of a network preserves its value independently of the considered random sub-sample [51].

We use the Jaccard coefficient to estimate the similarity of the neighbourhoods of two adjacent nodes, that is to say the ego networks of two socially tied individuals. This is a very important index, as it describes the correlation between different ego networks. Capturing this aspect is one of the key goals of our model. The Jaccard coefficient for two sets A and B is defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ and it is also not biased by random sub sampling[3]. Since computing the Jaccard coefficient between the end-points of each social link in the network requires huge computational efforts, we estimate its average value considering the pairs of end-points of a sample of $10,000$ edges randomly extracted from the network. The estimated average Jaccard coefficient (global) is reported in Table 4.5 (95% confidence level is indicated between square brackets). According to this result, considering two socially connected individuals, their common acquaintances are, on average, 4% of the union of their acquaintances. Intuitively, individuals connected by strong ties should have a higher ego network similarity than individuals connected by weak ties. In order to verify this intuition, we sampled $10,000$ edges for each circle of the ego networks (external part) and computed the Jaccard coefficient between the ego networks of the nodes at the endpoints of the links. As expected, results, reported in Table 4.5, confirm that the similarity is higher for inner circles and lower

[3] This can be easily seen observing that random sampling proportionally affects both the union and the intersection sets.

Table 4.5: Structural properties of the reference and generated networks.

	reference network	$p = 0.8$ $d_{min} = \frac{250}{n}$	$p = 0.8$ $d_{min} = \frac{500}{n}$	$p = 0.8$ $d_{min} = \frac{1{,}000}{n}$	$p = 0.5$ $d_{min} = \frac{500}{n}$
mean degree	27.82	133.91	133.94	134.00	133.86
avg. shortest path	4.06	3.40	3.26	3.11	3.12
clustering coefficient	0.109	0.152	0.108	0.085	0.079
Jaccard (global)	0.038 [±0.001]	0.060 [0.001]	0.040 [±0.001]	0.030 [±0.001]	0.030 [±0.000]
Jaccard (support cl.)	0.069 [±0.001]	0.084 [±0.001]	0.071 [±0.001]	0.064 [±0.001]	0.042 [±0.001]
Jaccard (sympathy gr.)	0.056 [±0.001]	0.073 [±0.001]	0.059 [±0.001]	0.053 [±0.001]	0.036 [±0.000]
Jaccard (affinity gr.)	0.042 [±0.001]	-	-	-	-
Jaccard (active net.)	0.031 [±0.001]	0.059 [±0.001]	0.037 [±0.000]	0.025 [±0.000]	0.030 [±0.000]

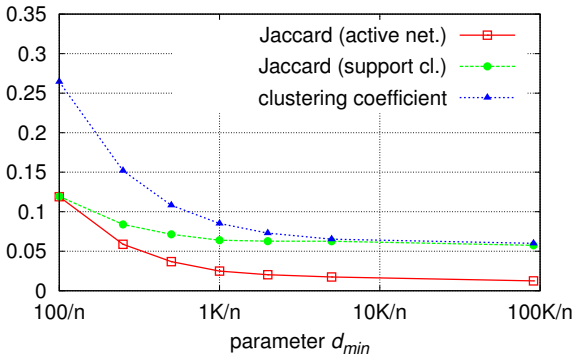

Figure 4.9: Clustering coefficient and Jaccard indexes for different d_{min} (with $p = 0.8$).

for outer circles. Specifically, it drops from about 7% for the support clique to about 3% for the active network.

4.2.5 Results and Validation

The majority of the parameters for the model described in Section 4.2.2 are directly inferred from the socio anthropology literature as discussed in Section 2.3. The only parameters we can set in order to conduct experiments are: (i) the number of nodes in the network n; (ii) the probability of selecting the "triadic closure" strategy, and (iii) the minimum distance d_{min} for f_D. In our experiments we choose to set $n = 91,347$, which is the number of nodes in the reference network, while we use different values for the parameters p and d_{min}. The main properties of the generated network are reported in Table 4.5. Note that, as for the single-ego model, generated networks do not consider the presence of the "affinity group" circle which we can assume to be merged with the "active network" circle.

The values of the parameters that allow us to best match the properties of the reference networks are $p = .8$ and $d_{min} = 500/n$ (fourth column of the table). These values mean that 80% of the social relationships are established through the triadic closure mechanism, rather than creating a bridge, and that, given a node, the 500 closest nodes (on average) have the same highest-probability to be selected as link's destination. Results show a strikingly similarity of the social structures between the reference network and the graph generated though the model. Indeed, both networks have the same clustering coefficient and similar Jaccard indexes for the different ego network circles. Note that discrepancies in the

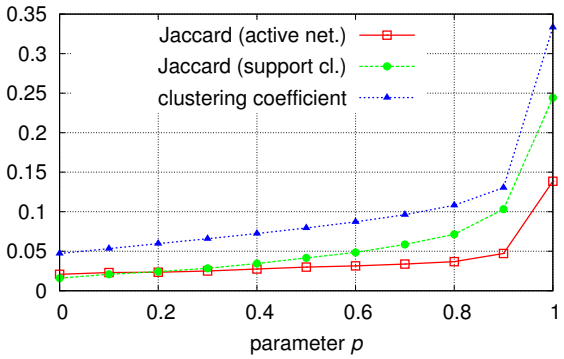

Figure 4.10: Clustering coefficient and Jaccard indexes for different p (with $d_{min} = 500/n$).

mean degrees and in the average shortest path length are due to the sub-sampling of the reference network. Remember that apart from these results for the global network, the use of the single-ego model (see Section 4.1) guarantees that well-known ego network properties are also satisfied. They are the size distribution of the network and of the single circles, the correlation between the circle dimensions and the distribution of the emotional closeness level.

In Table 4.5 we report the properties of the networks obtained with $d_{min} = 250/n$ (third column of the table) and $d_{min} = 1,000/n$ (fifth column of the table), maintaining $p = .8$. Moreover, Figure 4.9 shows the clustering coefficient and the Jaccard index computed between pairs of strongly-tied egos (i.e. belonging to each other support clique) and weekly-tied egos (belonging to each other active network). Results show that reducing d_{min} the clustering coefficient and the similarity indexes increase for all circles of the network. Intuitively, this is because with smaller d_{min} the set of nodes selected with highest probability by an ego (those at a maximum distance of d_{min}) is smaller, and geographically very close to the ego. This leads to higher clustering (and similarity).

Similarly to the geographical constraints, also the variation of the parameter p influences the structure of the network. As shown in the last column of the table and in Figure 4.10, if we diminish the value of p, the clustering coefficient and the similarity indexes decrease. This is expected as the number of links established though the bridging increases, and the bridging mechanism alone leads to the generation of random networks without clusters of socially connected nodes. Note in particular that when $p = 0$ (corresponding to a network without triadic closures)

the Jaccard indices in Figure 4.10 are the same, as in a network without triadic closures the correlation between social links do not depend on the strength of the links any more.

4.3 Discussion

In this chapter we have presented two unifying constructive network models for generating synthetic human social networks, in the perspective of Future Internet social networking environments. The aim of the first model, called single-ego model, is the generation of independent ego networks that match the structural properties highlighted in the social anthropology literature. To overcome the inherent limitation of the single-ego model, we have defined the multi-ego model, that significantly extends the first. It introduces different strategies to combine ego networks in order to generate complete social network graphs. The ego networks integration strategy is based on well-known properties in the field of social networks analysis i.e. (i) the triadic closure, (ii) the presence of bridges and (iii) the geographical constraints.

In order to validate our models, we have taken advantage of the convergence between the physical and the virtual worlds, thus using a real large scale human network obtained from Facebook. Tuning the model parameters we have obtained graphs with the same structural properties of the reference network. Then, we have analysed the effect of key parameters on the properties of the generated graphs, highlighting the impact of both geographical constraints and social constraints. The results confirm that our models lead to the generation of synthetic human networks that are consistent with the human social behaviour described in socio anthropology literature.

To the best of our knowledge, providing such unifying models is an original contribution of this thesis. These can be used both for the analysis, through large scale simulation, of key properties of human social networks and for the development of Future Internet solutions, that are going to be characterised by a strong link between the properties of the physical world and those of the cyber world. Being grounded in well established results from the social anthropology domain, our constructive model of human social network can be a very useful tool to characterise the properties of networking solutions in these Future Internet environments.

5

Modelling Information Diffusion in OSNs

Hitherto, we have focused on the analysis and modelling of social networks, highlighting which of their characteristics can be profitably used by researchers for the design of Future Internet solutions, e.g. the strength of the social links, the layered structure of the ego networks and the macro-level properties. Now we investigate the role of the network properties on a the diffusion of information in social network that is an important dynamic social phenomenon. In fact, understanding how information spreads between people could provide important insights into the dynamics of our society, revealing how the spread of ideas, innovation, influence and many other aspect take place. The advent of OSNs allowed scientists to remarkably improve the knowledge of the mechanisms controlling the formation of information diffusion chains in social networks (typically referred to as information cascades), however the role of many factors related to the human social behaviour still need more in-depth investigation.

In order to contribute to the characterisation of the information diffusion in OSNs we analyse the role of the users' activity in Twitter. For this reason we define an agent-based model to reproduce the behaviour of the users, such that the impact of the various parameters on information diffusion can be studied "in vitro". For example, one of the most important factors for the formation of information cascades is the *decaying visibility* of the content. In fact, different studies have demonstrated that the probability that a user forwards a received content decreases with time [31, 42, 50]. We believe that, for a better characterisation of the content visibility, it can not be measured only in terms of time and that the users' activity patterns should be considered too.

Focusing on Twitter, a more straightforward way for estimating the visibility of a *tweet* is considering its *position* in the tweet feed that is the result of the global

users' activity. In fact, as empirically demonstrated in Section 5.3.1, the tweet's position in the feed is strongly correlated with its probability to be retweeted giving rise to *information cascades*. In addition to the position of a tweet in the user's feed, we also show that other parameters related to the user that originally generates a tweet can impact on the diffusion of information in Twitter. We collectively represent them with a unique parameter, that we call *user standing*. These properties are the base for the agent-based model we describe in Section 5.4. In the model, agents simulate the users' activity in creating new messages and forwarding previously received messages. Basing on an underlying network structure, messages are dispatched to the connected agents and, based on their position in the tweet feed and the standing of the originating agent, they are probabilistically forwarded, simulating the formation of information cascades. We evaluate our model (simulating the user activity) in a network whose parameters are derived from a large Twitter data set we present in Section 5.1). Simulation results match empirical observations with high statistical confidence both in terms of information cascade properties and characterisation of the user influence.

Moreover, the fact that the information diffusion is driven by the activity of the agents, makes the model suitable for a wide range of applications. For example, assuming the spread of information results unfair because of the disproportional influence of the users with a large number of followers, we simulate the effect of an hypothetical mechanism for limiting this inequality. It consists in reducing the visibility of certain messages modifying their position in the message feeds. Simulation results show that turning down some messages of just 10 positions is enough to balance the user influence in the network significantly.

5.1 Data Set Description

For the analysis of the user behaviour in OSNs we downloaded a large data set of user activity from Twitter that is one of the most important social networking service.

5.1.1 Platform Description

Twitter is an online social networking and microblogging service founded in 2006, with more than 500 million registered users as of 2012[1]. In Twitter, users can post short public messages (with at most 140 characters) called *tweets*. All the users'

[1] According to Twitter CEO Dick Costolo in October 2012.

tweets are accessible by other users, unless the users' profiles are private or the access is restricted by other specific settings. Users can also automatically receive notifications of new tweets created by other users by "following" them (i.e. creating a subscription to their notifications). People following a specific user are called her *followers*, whilst the set of people followed by the user are her *friends*.

Tweets can be enriched with multimedia content (i.e. URLs, videos, pictures) and by using special text characters to insert additional information. Specifically, a tweet can reference one or more users with a special mark called *mention*. Users mentioned in a tweet automatically receive a notification, even though they are not followers of the tweet's author. Users can also *reply* to tweets. In this case, a tweet is generated with an implicit mention to the author of the replied tweet. Replies often involve bi-directional communications, since they are usually used to reply to previously received mentions. Twitter has also a private messaging system, however, since private messages are not publicly accessible, we did not collected them in our data set.

In addition to mentions and replies, Twitter provides a series of mechanisms for broadcast communication that represent the most popular features of the platform. Firstly, all the tweets are automatically sent towards all the followers of their authors. Moreover, tweets can also be *retweeted*. A user can make a retweet to forward a tweet it to all her followers. Each tweet can be assigned to a topic through the use of a special character called hashtag (i.e. "#") placed before the text indicating the topic. Hashtags are used by Twitter to classify the tweets and to obtain *trending topics*.

5.1.2 Data Download

We implemented a crawling agent which is able to download user profiles and their communication data from Twitter. The agent visited the Twitter graph considering the users as nodes and following the links between them. In particular, we assume that a link between two nodes exists if at least one of the users follows the other or an interaction between them has occurred. We use as indication of an interaction the presence of a *mention* in a tweet (i.e. the fact that a user explicitly mentions the other in a tweet) and a *reply* (i.e. a direct response to a tweet).

The crawling agent starts from a given user profile (seed) and visits the Twitter graph following the links. For each visited node, we took advantage of the Twitter REST API to extract the user *timeline* (i.e. the list of posted tweets that can include mentions and replies), the *friends* list (i.e. the people followed by the user) and the *followers* list (i.e. the people who follow the user). Twitter REST API limits the

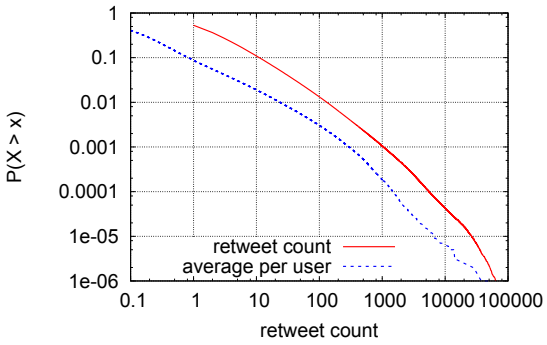

Figure 5.1: CCDFs of retweet count and average retweet count per user (influence).

amount of tweets that can be downloaded per user to the last $3,200$ tweets. This does not represent a constraint to our analysis since, as we show in the following, it is sufficient for our purposes.

The crawling agent used 250 threads that concurrently accessed a single queue containing the ids of the user profiles to download. Each thread extracted a certain number of user ids from the queue, then it got the related profiles and communication data from Twitter using the REST API. Finally, after extracting new user ids from the communication data and from the friends/follower lists, the threads add them to the queue. The use of multiple threads allowed both to speed-up the data collection and to avoid the crawler to remain trapped in visiting the neighbourhood of a node with a large number of links. The seed we used to start the data collection is the profile of a widely know user (user id: 813286), so that her followers represent an almost random sample of the network.

The crawling agent has been active from November 2012 to February 2013, collecting the data of $2,029,143$ Twitter users. In total the data set contains around $2,500$M tweets that we divided in *"regular" tweets* (63.2%), *replies* (19.9%) and *retweets* (16.9%). Replies are often related to personal conversation and they have not an active role in the propagation of information in the network. In fact, analysing the retweets in our data set, we discovered that only 1.04% of them are related to replies. For this reason, in our analysis we ignore the replies tweet and we consider just "regular" tweets and their retweets.

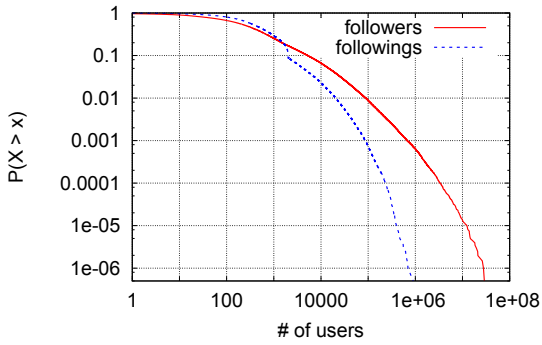

Figure 5.2: CCDFs of number of followers and followings.

5.2 Influence in Twitter

The influence can be defined as the ability of a user to spread information in a network. In the literature several measures of influence have been proposed. Some of them consider the structural features of the nodes such as the number of followers [7, 91, 50], the PageRank [50, 88], and various centrality measures [88, 11]. Other measures, considered more reliable, use instead the past network activity quantifying for each user the effective propagation of his messages [7, 50, 12, 52]. In Twitter, the propagation of a message can be measured in terms of *retweet count*, that is the number of times the message has been retweeted and that is included in the metadata of each downloaded tweet. Using this information we can define the influence of a user in Twitter as the average retweet count of all tweets he created.

Figure 5.1 displays the Complementary Cumulative Distribution Functions (CCDFs) of the retweet count and of the user influence by the solid and dotted lines respectively. These results are inline with other analysis in literature that have shown that the size of information cascades and the user influence tend to be highly skewed [31, 7, 12]. This means that only a small fraction of users can be considered influential. In fact, in our data set, only about 0.01% of them are able to get their tweets to be retweeted more than $1,000$ times on average.

Starting from the measure of influence, we can now examine what factors are related to it using our data set. Literature says that the structural feature that best correlates with the user influence is the number of followers [7, 50, 76] that corresponds to the in-degree of the nodes in the underlying network topology. The reason behind is that a tweet from a user with many followers reaches immediately a

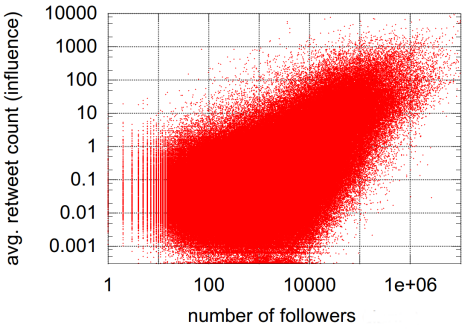

Figure 5.3: Relation between # of followers and influence.

large audience that, possibly, will retweet it to other users. In Figure 5.2, we show the CCDFs of the number of followers (in-degree) and followings (out-degree) by the solid and dotted line respectively. Like any other social network, the degree distribution of the Twitter graph has a long tail. This means that, compared with the total number of users, just a small fraction of them have a very large audience and, as suggested in [50] and [12], they are probably celebrities and mass media.

In Figure 5.3 we show the log-log plot of the number of followers against the user influence. The correlation (Pearson coefficient equal to 0.532) is remarkable, however, given the same number of followers, the influence value can vary significantly. In fact, as previously mentioned, structural features of the nodes alone are not sufficient to explain the actual influence of a user in the network. Others factors should be investigated.

5.3 Factors on Retweeting Behaviour

Having discussed the influence of the number of followers on the propagation of the messages from a global network standpoint, we analyse what leads individual users to retweet messages they find in their tweet feeds. In other words, we try to highlight the factors that determine the retweeting mechanism at the user level, that collectively generate the overall effect highlighted in the previous section. When a Twitter user accesses his tweet feed there are different factors that impact on his behaviour leading him to select a message to retweet. We perform our study by assuming that two main factors impact on the detailed retweeting behaviour of the users: the position of tweets in the feed, and an overall parameter describing all the properties of the creator of the tweet, that we call *user standing*.

In principle, the content of tweets also may have a role. However, analysing this would require detailed analysis on the semantic of tweets and on the interests of users. To keep the analysis simple we don't consider these aspects, and assume that the user standing also captures the average quality of the tweets' content.

5.3.1 Position in the Tweet Feed

Previous studies have inferred that visibility of the tweets is related to their probability to be retweeted [31, 42, 50]. A tweet has the maximum visibility immediately after it is received because it takes the least effort to be discovered at the top of the tweet feed. As soon as new tweets arrive in the feed, they push the old messages down in the queue reducing their visibility. Oken Hodas and Learman [42] have also noted that this effect is more dramatic when a user follows more people.

We believe that the time span after receiving a tweet is a good estimator of its visibility however, it can be influenced by other factors like the temporal activity patterns of the users. A more straightforward approach, is to analyse the actual position of the messages in the tweet feed. For this analysis we randomly selected a subset of $100,000$ users from our data set. Then for each user we have recreated his message feed joining all the published tweets of the users he follows. Successively, comparing the timestamps, we have extracted for each retweeted message its position in the tweet feed at the time of the retweet. In our analysis we have considered only the first $1,000$ positions of the feed. The retweet of messages beyond this threshold could reveal a non-typical approach of the user who, for example, should have read the message accessing directly to a profile page rather than scrolling his tweet feed. Results in Figure 5.4 show that the probability of retweetting a message in a certain position of the feed follows a power-law distribution with coefficient 1.433 estimated using the maximum-likelihood estimation (MLE).

It is worth noting that the position of the messages in a tweet feed is pretty much random, since it depends only on the time a user receives the messages and on the time he retweets. The relation between the position and the retweet probability, therefore, does not explain the variation on the user influence discussed at the end of Section 5.2. Visibility is, in fact, a general property of the tweets and doesn't depend on the influence or on the number of followers of the users.

5.3.2 User Standing

In order to explain mentioned variations in the user influence we have to investigate the effect of the properties of the users on the retweeting behaviour. These proper-

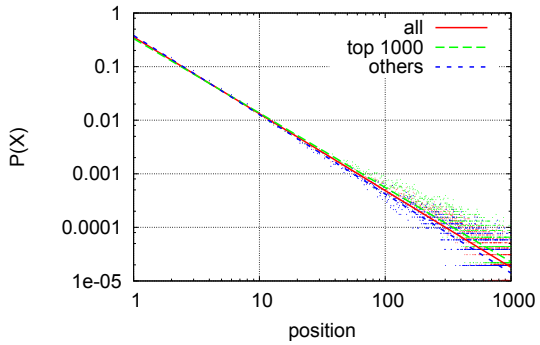

Figure 5.4: Retweet probability given the position in the feed for all the tweets in the data set ("all"), for the tweets created by the $1,000$ most influential users ("top 1000") and for the tweets created by all the other users ("others").

ties are often qualitative and, therefore, hard to quantify (e.g. credibility, expertise, enthusiasm and popularity). For this reason, we use a unique index called *user standing*, to take into consideration the joint effect of all of them.

The effect of the user standing can be observed as the variation of the retweet probability for different equally-positioned tweets. In this sense, the user standing can be considered as a sort of "favouritism in retweet" for the messages created by some users. In our case, we are interested in investigating if the tweets created by the most influential users are more likely to be retweeted than the tweets created by other users. For the analysis we ranked the users considered in the previous analysis based on their influence and then we selected the top $1,000$ influential users. In Figure 5.4 we plot the retweet probability of their tweets compared with the retweet probability of tweets created by all the other users. The gap between the lines appears narrow, however the fit with a power-law function has coefficient 1.389 for the influential users and 1.478 for the others. This means that, considering the same position, the most influential users have a higher probability to get their messages retweeted.

5.4 Activity-Based Propagation Model

The model we present in this work describes the information propagation mechanism in a microblogging social network given the topology of the network and some features of the agents that represent the users. In the model any agent

interacts with the network in two different ways: creating new messages and forwarding previously received messages. The frequency with which an agent v is selected for creating and forwarding messages, is given by the parameters f_v^{cr} and f_v^{fw} respectively. Both in case of creation and forwarding, the messages are broadcast to other agents that "follow" the creator or forwarder. An agent r follows the agent v if, in the underlying network graph $G(V, E)$, a direct link between the nodes that represent agents r and v respectively exists. In this case the agent r receives all the messages created or forwarded by agent v. If an agent receives multiple copies of the same message, it keeps in memory just the first received one and discharges the others.

Assuming that an agent v is selected to perform a forwarding action at time t, the model takes the *message feed* list $F_{v,t}$ that includes all the messages received by v before time t sorted by reverse-chronological order. Then, for each message $w \in F_{v,t}$, it assigns the probability $P(w|v,t)$ to be forwarded such that $\sum_{w \in F_{v,t}} P(w|F_{v,t}) = 1$ where:

$$P(w|v,t) = \frac{\alpha_{o(w)}\varphi(\theta_{v,t}(w))}{\sum_{z \in F_{v,t}} \alpha_{o(z)}\varphi(\theta_{v,t}(z))}, \qquad w \in F_{v,t} \tag{5.1}$$

$\alpha_{o(w)}$ is the *standing* of the the agent $o(w)$, who is the creator of the message w, and $\varphi(\cdot)$ is a function called *position function* that takes as a parameter the position of w in $F_{v,t}$ denoted as $\theta_{v,t}(w)$. According to Equation 5.1, the probability of a message to be selected for the forward depends on: i) its position in the message feed and ii) the standing of its creator.

i) The position of the message in the feed is considered in the model since, as we demonstrated in Section 5.3, there is evidence that last received messages (which are on top of the message feed) are more likely to be forwarded. For this reason the position function $\varphi(\cdot)$ has to be monotonically decreasing. For example, as our analysis suggests, it can be defined as a power-law function.
ii) As discussed in Section 5.2, we introduced the concept of user standing that represents the joint effect of all the properties of the users that positively influence the forwarding probability of their messages. Each agent in the network v is therefore characterised, in addition to the frequencies f_v^{cr} and f_v^{fw}, also by a standing value α_v. In the next section we discuss in detail how to model the user standing.

Table 5.1: Social graph statistics.

#nodes	$100,000$
#arcs	$5,756,450$
mean degree	57.565
clustering coefficient	0.156
average path length	3.557
diameter	14

5.5 Deriving the Model's Parameters

In our simulation we implement the agent-based propagation model described in previous section in order to simulate the user activity and the information diffusion of a real social network. We used the Twitter data set described in Section 5.1 to infer both the graph structure and the agents' properties.

5.5.1 Social graph

For computational reasons we selected a random subset of $100,000$ users among all the active users from our data set. We considered a user to be active if he has at least 100 followers and if he has created at least 100 tweets. These constraints allow us to avoid low-active accounts that are not relevant for the propagation of information. From this set of users, we derived the social graph whose relevant statistics are summarised in Table 5.1. The social graph maintains well-known features of social networks' graphs such as high clustering coefficient and small average path length (small-world property) [64].

5.5.2 Position Function

As suggested in Section 5.4 we define the position function $\varphi(\cdot)$ as a power-law. In particular we use the result in Section 5.3 in which we have fit the retweet probability given the tweets' position with a power-law with coefficient 1.433. Considering that the position function is discrete, we define it as a ZipF Probability Mass Function with the given coefficient and limited to $N = 1,000$, which is the same number of positions we have used in our analysis.

5.5.3 Frequencies

For each user v we extract, from the data set, the frequency of creating messages per day f_v^{cr} and the frequency of forwarding messages per day f_v^{fw}. Distributions

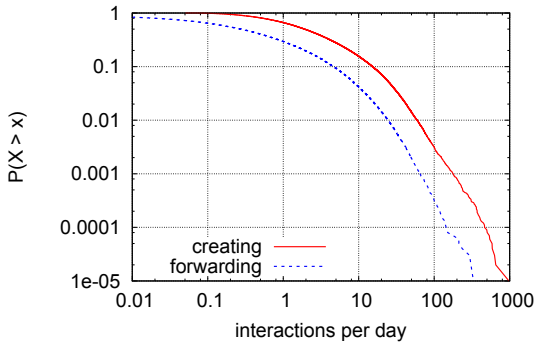

Figure 5.5: CCDFs of the frequencies of interaction.

of these frequencies, shown in Figure 5.5, highly skewed since just few users have a very high activity.

5.5.4 User Standing

In Section 5.3.2, we defined the user standing as the joint effect of the latent factors that affect the forwarding of his messages. As previously discussed, these parameters of the model are not directly quantifiable. We could estimate them using a MLE estimator where the likelihood function is given by a sample of retweeting actions extracted from the data set. Unfortunately, applying this method would have required to analyse the full propagation path of each and every tweet of all our users, which was not feasible due to the computational complexity and the fact that cascades can involve users not included in our data set. Therefore, we use an approximate way to estimate the user standing, as follows.

The idea is to estimate the standing of a user as the average retweet probability of the tweets he has originated. This can be calculated as the ratio of his average retweet count (influence) to the average number of users who have received his tweets. However, the latter value is not derivable since it would require to track the full propagation trees. As approximation, we use the number of his followers instead. It is worth noting that, due to this approximation, the standing of the most influential users could be overvalued. This is because the number of followers can be significantly smaller than the number of users that received the tweets. In order to remove this bias we had to apply an exponent to the previously defined measure. As result of an extensive analysis, we set the exponent to $1/3$ as this

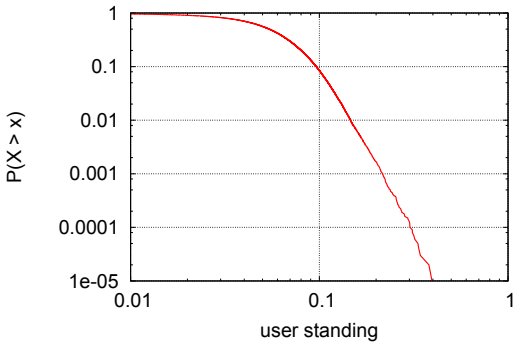

Figure 5.6: CCDF of user standing.

value guarantees to obtain better performance of our model. Formally, the user standing values we considered in our simulation are defined as:

$$\alpha_r = \left(\frac{\sum_{w \in W_r} \pi(w)}{|W_r| \cdot k(r)} \right)^{1/3} \tag{5.2}$$

where w is a message, W_r is the set of messages created by user r, $\pi(w)$ is the number of times the message w has been forwarded and $k(r)$ is the number of followers of the node r. The CCDF of the obtained values is shown in Figure 5.6.

5.6 Simulations

Using the social graph and the user parameters described in Section 5.5, we simulated a period of 30 days of user activity. We run 10 independent simulations in order to calculate the 95% confidence intervals which are shown as error bars in the figures and between square brackets in the tables and in numerical data. The simulations produced an average of $24,026,886$ $[\pm 292]$ user interactions in that 77.1% $(18,515,225$ $[\pm 1,092])$ are related to the creation of new messages and the rest are forwarding messages. These proportions are consistent with those related to the data set in Section 5.1 (excluding reply tweets). Among all created messages, 14.3% of them $(2,649,709$ $[\pm 1,128])$ have been forwarded originating cascades. In Figure 5.7 we show the histogram of the depth of the cascades produced. As we can see, the trend is logarithmically decreasing with respect to the frequency. In fact, 78.7% of the forwarded messages are not propagated beyond

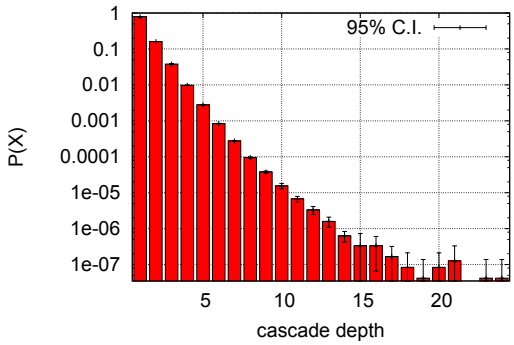

Figure 5.7: Cascade depth distribution.

the first level of followers. This trend is exactly the same shown in several analysis in literature [31, 7].

As discussed in Section 5.2, we define, for each node r in the simulations, the influence γ_r as the average retweet count of the tweets r has originated. In Figure 5.8 we show the CCDFs of the number of forwards for each message as the solid line and the nodes' influence as the dashed line. Comparing these results with those in Figure 5.1, we can see that the simulations replicated the presence of a small number of influential users located in the tail of the distribution[2].

In the column "orig" of Table 5.2, we summarise the results of the simulations (upper part) and the correlation of the resulting influence with other variables (lower part). In the table we refer to the the vector of the nodes' influence as γ while we use the symbol $\boldsymbol{k}$ for the vector of the number of followers and α for the vector of the users' standing. Correlation values demonstrate that our model is able to replicate high correlation between the influence and both the number of followers and the user standing [3]. We also calculate the correlation between the simulated user influence and the influence γ* of the selected users in the data set described in Section 5.1. Considering that the influence from the data set refers to the actual influence of the users in the Twitter network and that in our simulations we consider just a small subset of this network, the correlation value is remarkable and proves the ability of our model to simulate the actual user influence distribution.

[2] Direct comparison between the two plots is not possible, due to the large difference of the number of users in the data set and in the simulations.

[3] Note that, while in Equation 5.2 the standing is clearly a function of the influence, the values of the user standing have been computed based on the information propagation in the data set, while influence is measured based on the simulations' results.

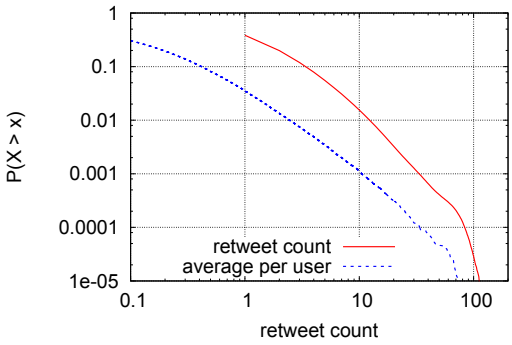

Figure 5.8: CCDFs of forwardings per message and user influence.

5.7 Message Positioning and User Standing Impact

In order to study the impact of the message positioning in our model we run 10 simulations with the same setting described in Section 5.5, excluding the position function $\varphi(\cdot)$ from the model. Results are shown in the column "no-pos" of Table 5.2. The main consequence of such change is that some messages flood the entire network and some users become extremely influential. This indicates that the decreasing visibility of the messages in social networks is fundamental for limiting the size of the information cascades.

We also studied the impact of the user standing, running 10 simulations where we have excluded it from the probability of forwarding. In this case the main change in results, shown in column "no-sta" of Table 5.2, is an increase in the correlation between the number of followers and a decrease in the correlation between the influence and the standing values.

In both "no-pos" and "no-sta" cases, it is noticeable the sensible decrease of the correlation between the simulated influence and the actual influence registered in the our data set. This demonstrates the importance of considering both parameters in our model.

5.8 A Case of Study: Smoothing Users Influence

As we discussed in Section 5.2, in Twitter the users with a large number of followers tend to become the most influential, based on the fact that they can easily reach other users due to their high degree. This is also captured in our model, as

Table 5.2: Summary of the simulation results.

	orig	no-pos	no-sta
cascade depth	19.0 [± 2.1]	121.6 [± 6.0]	10.0 [± 0.5]
max msg forwards	257.7 [± 37.2]	$10,347.4$ [± 304.2]	155.9 [± 7.0]
max user influence	79.1 [± 0.6] [id:41801]	$1,436.0$ [± 187.6] [id:98020]	131.6 [± 5.2] [id:2019]
corr(γ,$\boldsymbol{k}$)	0.544 [± 0.010]	0.100 [± 0.009]	0.646 [± 0.003]
corr(γ,$\boldsymbol{\alpha}$)	0.101 [± 0.003]	0.073 [± 0.004]	0.014 [± 0.001]
corr(γ,γ^*)	0.595 [± 0.003]	0.126 [± 0.011]]	0.443 [± 0.003]

shown by the correlation between the resulting influence and the number of followers (see Table 5.2). This disproportional influence may be perceived as generating unfairness in the network, inhibiting opinions from other users to come out. Assuming the need of reducing this lack of fairness, we can simulate the consequence of introducing a mechanism that penalises the users with high degree modifying the visibility of their messages.

We run 10 simulations in which we modify the message feeds ordering, lowering the position of the messages created by the users that have at least $1,000$ followers. The number of assigned penalty positions varies between one position, for nodes with $1,000$ followers, and a maximum of 10 positions for the user with maximum degree ($14,653$ followers).

In Table 5.3 we compare the attributes of the top 10 influencers obtained using the original model (column "orig") and using the model with penalties (column "penalty"). As we can see, some of the users with high degree are still in the list but their influence has been strongly reduced. For example, the user with maximum degree (the node 41801) reduced his influence from 70.15 to 36.73. On the other hand, many users with low degree entered in the ranking.

5.9 Discussions

We analysed the properties of the information diffusion in Twitter, in particular the impact of the structural features of the users and their retweeting behaviour. Using a Twitter data set we studied the relation between the probability of a message to be retweeted and its position in the tweet feed and we concluded that this relation is described by a power-law function. We also characterised the most influential

Table 5.3: Top 10 most influential users in simulations using the original model (column "orig") and the penalty model (column "penalty").

	orig				penalty			
#	id	k	α	γ	id	k	α	γ
1	41801	4,679	0.079	78.88 [±0.68]	2019	11,368	0.016	45.00 [±1.61]
2	2019	11,368	0.016	70.15 [±4.61]	41801	4,679	0.079	36.73 [±0.31]
3	5201	14,654	0.037	67.73 [±0.67]	64132	1,301	0.104	34.80 [±3.48]
4	8187	6,429	0.023	60.70 [±1.23]	8187	6,429	0.023	34.08 [±1.58]
5	4619	3,555	0.086	55.07 [±2.72]	4619	3,555	0.086	32.08 [±2.15]
6	23317	4,280	0.049	45.01 [±0.76]	88180	572	0.180	32.04 [±1.55]
7	64132	1,301	0.104	42.66 [±1.12]	91505	217	0.274	28.68 [±3.35]
8	3617	7,938	0.053	40.88 [±1.01]	83024	518	0.166	26.71 [±2.10]
9	15145	6,993	0.036	38.59 [±0.48]	32210	651	0.099	25.80 [±1.05]
10	89312	2,140	0.065	34.26 [±0.86]	89312	2,140	0.065	25.04 [±0.70]

users in the network discovering that, although their ability of spreading messages is mostly given by their large number of followers, other factors have to be considered. These factors, joint effect we called user standing, have effect at the forwarding behaviour level, scaling the retweet probability given by the position of the message.

Based on these observations we proposed an agent-based information propagation model able to generate cascades whose properties match empirical observations. Agents simulate the activity of the users in a network creating and forwarding messages independently. Received messages are organised in an ordered list for reproducing the effect of the position on the forward probability.

Through simulations, we show that our model is able to reproduce information cascades statistically similar those presented in the literature and that the generated user influence is strongly correlated with the actual influence measured in the data set. These results demonstrated that our model can thus be used to realistically study how the user activity and the forwarding mechanism influence the propagation of information.

As a case of study, we simulated the introduction of a strategy for smoothing the influence of users in order to make the information diffusion more fair. Specifically, we modified the message feed ordering for limiting the influence of users that have large number of followers.

6

Conclusions

In this thesis we have first verified the convergence between the physical and the cyber worlds though an extensive analysis of the structural properties of the social networks in both the environments. Then, we have exploited this results to define new models of social networks and information diffusion that can be profitably used for the design and testing of Future Internet solutions.

As far as the cyber-physical convergence is concerned, we have started discussing some of the most important results in the fields of psychology and socio anthropology about the human social networks. Specifically, we have focused on the impact of the cognitive limits of the human brain on the way people maintain their social relationships, leading to the formation of typical social network structures called ego networks. In order to better characterise these structures, we have placed emphasis on the definition of the strength of the social relationships. In fact, variations on the level of the tie strength allowed researchers to describe the properties of the different layers (called circles) into which an individual organises the active social relationships that belong to his/her ego network.

A fundamental contribution of this thesis is given by the first extensive study on the similarities between the structures observed in human social networks and the properties of OSNs, formed by ICT users in cyber environments. To this aim, we analysed a large data sets of social traces obtained from Facebook. Given data set has been processed in order to discharge inactive and non-representative users and to estimate the strength of the social links as a function of the frequency of contact, as suggested by results in the reference literature. Finally, we have isolated a considerable amount of online ego networks whose layered structure has been extracted through an accurate clustering analysis.

Comparing the results of this analysis with the known properties of the human social networks, we have demonstrated that the ego networks observed in physical and cyber environments are strictly similar. In particular, we have verified that around 75% of the Facebook ego networks exhibit a typical number of social circles equal to four, that is the same number of circles into which humans organise their social relationships in the physical world. Another important evidence of the convergence between human and online ego networks is the scaling factor between the size of the social circles. Indeed, we have observed an average scaling factor of 3.12, that is very close to the reference value of 3.06, suggested by the literature about human ego networks. These results has been further verified by a similar analysis that use a data set from Twitter.

Basing on the convergence between the physical and the cyber worlds, we have designed two generative network models including the results obtained in the human social network domain. The original element of proposed models is given by a characterisation of the social links at a higher level of detail with respect to other available solutions. In fact, through using insights on human social relationships, our models assign a certain level of tie strength to each social link, so that generated ego networks are consistent with the properties observed in physical environments. Furthermore, ego networks are conveniently integrated in order to satisfy well-known macroscopic features of the social networks, for instance the small-world property and the presence of geographical constraints.

Our models have been validated comparing generated networks with a large social graph obtained from Facebook. The analysis verified that, in addition to the properties on the ego networks, generated networks are compatible with real social graphs also in terms of average shortest path, clustering coefficient and Jaccard similarity index. Proposed models can be profitably used by researchers for the design of Future Internet solutions. In fact, the strength of the social links can be used as a proxy of the frequency of contact between people, playing a fundamental role in the design of networking solution for efficient content dissemination in electronic networks. Moreover, generated graphs can be used for the developement and testing of advanced social services taking advantage of the detailed characterisation of the social links.

In the perspective of such an integrated cyber-physical world, virtual communities play a fundamental role on the creation and propagation of content. For this reason we have also investigated the information diffusion phenomenon in OSNs. Specifically, we have analysed the social behaviour of the users in Twitter and the role of the content visibility in the diffusion of information. Through the use of a specific methodology, we have measured the visibility of a content in terms of the

position of the message in the user's tweet feed. To the best of our knowledge, this is the most accurate measure of the content visibility in Twitter since, in other studies, it was estimated through the lifetime of the message which is less reliable since it can be influence by other factors as the number of friends and their activity. Analysing the position of retweeted message at the time of the retweet, we observed that its distribution can be fitted using a power-law function with a coefficient between 1.389, for the most influential users, and 1.478, for the others. This discrepancy suggests that the influence of the users is also affected by latent factors whose investigation represents an interesting future work.

Based on this analysis we designed a model of information diffusion that accurately reproduces the behaviour of the users in Twitter and that can be used for the design and testing of advanced social networking services.

References

1. Richard D Alba and Charles Kadushin. The intersection of social circles a new measure of social proximity in networks. *Sociological Methods & Research*, 5(1):77–102, 1976.
2. Valerio Arnaboldi, Marco Conti, Massimiliano La Gala, Andrea Passarella, and Fabio Pezzoni. Information diffusion in osns:the impact of nodes' sociality. In *Proceedings of the 29th ACM Symposium On Applied Computing (ACM SAC 2014)*. ACM, 2014.
3. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin I M Dunbar. Dynamics of personal social relationships in online social networks: a study on twitter. In *Proceedings of the first ACM conference on Online social networks*, pages 15–26. ACM, 2013.
4. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of ego network structure in online social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 31–40. IEEE, 2012.
5. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Ego networks in twitter: an experimental analysis. In *INFOCOM, 2013 Proceedings IEEE*, pages 3459–3464. IEEE, 2013.
6. Valerio Arnaboldi, Andrea Passarella, Maurizio Tesconi, and Davide Gazzè. Towards a characterization of egocentric networks in online social networks. In *On the Move to Meaningful Internet Systems: OTM 2011 Workshops*, pages 524–533. Springer, 2011.
7. Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.
8. Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.
9. Lawrence E Blume. The statistical mechanics of strategic interaction. *Games and economic behavior*, 5(3):387–424, 1993.
10. Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1909–1912. ACM, 2010.

11. Rafael Cappelletti and Nishanth Sastry. Iarank: Ranking users on twitter in near real-time, based on their information amplification potential. *HUMAN JOURNAL*, 1(2):100–115, 2012.
12. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.
13. Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18. ACM, 2008.
14. Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.
15. Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.
16. Marco Conti. Special section on mobile opportunistic networking. *Pervasive and Mobile Computing*, 7(2):159, 2011.
17. Marco Conti, Song Chong, Serge Fdida, Weijia Jia, Holger Karl, Ying-Dar Lin, Petri Mähönen, Martin Maier, Refik Molva, Steve Uhlig, et al. Research challenges towards the future internet. *Computer Communications*, 34(18):2115–2134, 2011.
18. Marco Conti, Sajal K Das, Chatschik Bisdikian, Mohan Kumar, Lionel M Ni, Andrea Passarella, George Roussos, Gerhard Tröster, Gene Tsudik, and Franco Zambonelli. Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber-physical convergence. *Pervasive and Mobile Computing*, 8(1):2–21, 2012.
19. Marco Conti, Silvia Giordano, Martin May, and Andrea Passarella. From opportunistic networks to opportunistic computing. *Communications Magazine, IEEE*, 48(9):126–139, 2010.
20. Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model for the generation of social network graphs. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a*, pages 1–6. IEEE, 2011.
21. Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model to represent human social relationships in social network graphs. In *Social Informatics*, pages 174–187. Springer, 2012.
22. Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. An experimental study of search in global social networks. *science*, 301(5634):827–829, 2003.
23. Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.
24. Robin I M Dunbar. The social brain hypothesis. *brain*, 9:10, 1998.
25. Robin I M Dunbar. Theory of mind and the evolution of language. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight, editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge University Press, Cambridge, 1998.
26. Robin I M Dunbar and Matt Spoors. Social networks, support cliques, and kinship. *Human Nature*, 6(3):273–290, 1995.
27. David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1, 2010.

28. Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.
29. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
30. Noah Friedkin. A test of structural features of granovetter's strength of weak ties theory. *Social Networks*, 2(4):411–422, 1980.
31. Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, pages 3–3. USENIX Association, 2010.
32. Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.
33. Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.
34. Jacob Goldenberg, Barak Libai, and Eitan Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, 9(3):1–18, 2001.
35. Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.
36. Bruno Goncalves, Nicola Perra, and Alessandro Vespignani. Validation of dunbar's number in twitter conversations. *arXiv preprint arXiv:1105.5170*, 2011.
37. Mark S Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
38. Mark S Granovetter. *Getting a Job: A Study of Contacts and Carreers*. Harvard University Press, 1974.
39. Mark S Granovetter. Threshold models of collective behavior. *American journal of sociology*, 1978.
40. Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.
41. Russell A Hill and Robin I M Dunbar. Social network size in humans. *Human nature*, 14(1):53–72, 2003.
42. Nathan Oken Hodas and Kristina Lerman. How visibility and divided attention constrain social contagion. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 249–257. IEEE, 2012.
43. Bernardo A Huberman, Peter LT Pirolli, James E Pitkow, and Rajan M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.
44. Muhammad Usman Ilyas, Muhammad Zubair Shafiq, Alex X Liu, and Hayder Radha. A distributed and privacy preserving algorithm for identifying information hubs in social networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 561–565. IEEE, 2011.

45. Jason J Jones, Jaime E Settle, Robert M Bond, Christopher J Fariss, Cameron Marlow, and James H Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.
46. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.
47. David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.
48. Wouter Kool, Joseph T McGuire, Zev B Rosen, and Matthew M Botvinick. Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, 139(4):665, 2010.
49. Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.
50. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
51. Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.
52. David Leonhardt. A better way to measure twitter influence. *The New York Times*, 24, 2011.
53. Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010.
54. Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.
55. Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. *arXiv preprint arXiv:0704.2803*, 2007.
56. David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.
57. Nan Lin and Mary Dumin. Access to occupational resources through social ties. In *Annual Meeting of the American Sociological Association*, 1982.
58. Nan Lin, John C Vaughn, and Walter M Ensel. Social resources and occupational status attainment. *Social Forces*, 59(4):1163–1181, 1981.
59. Dunia López-Pintado. Diffusion in complex social networks. *Games and Economic Behavior*, 62(2):573–590, 2008.
60. Peter V Marsden and Karen E Campbell. Measuring Tie Strength. *Social Forces*, 63(2):482–501, 1984.
61. Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.
62. Stephen Morris. Contagion. *The Review of Economic Studies*, 67(1):57–78, 2000.
63. Stephen O Murray, Joseph H Rankin, and Dennis W Magill. Strong ties and job information. *Work and Occupations*, 8(1):119–136, 1981.

64. Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
65. Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.
66. Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4):e16939, 2011.
67. Andrea Passarella and Marco Conti. Characterising aggregate inter-contact times in heterogeneous opportunistic networks. In *NETWORKING 2011*, pages 301–313. Springer, 2011.
68. Andrea Passarella, Marco Conti, Chiara Boldrini, and Robin I M Dunbar. Modelling inter-contact times in social pervasive networks. In *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, pages 333–340. ACM, 2011.
69. Andrea Passarella, Robin I M Dunbar, Marco Conti, and Fabio Pezzoni. Ego network models for future internet social networking environments. *Computer Communications*, 35(18):2201–2217, 2012.
70. Subharthi Paul, Jianli Pan, and Raj Jain. Architectures for the future networks and the next generation internet: A survey. *Computer Communications*, 34(1):2–42, 2011.
71. Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002.
72. Sam GB Roberts and Robin I M Dunbar. Communication in social networks: Effects of kinship, network size, and emotional closeness. *Personal Relationships*, 18(3):439–452, 2011.
73. Sam GB Roberts, Robin I M Dunbar, Thomas V Pollet, and Toon Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2):138–146, 2009.
74. Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft. Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In *Proceedings of the 20th international conference on World wide web*, pages 457–466. ACM, 2011.
75. Leonard Schuchman. Dither signals and their effect on quantization noise. *Communication Technology, IEEE Transactions on*, 12(4):162–165, 1964.
76. Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184. IEEE, 2010.
77. Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M Lento. Gesundheit! modeling contagion through facebook news feed. In *ICWSM*, 2009.
78. Anjana Susarla, Jeong-Ha Oh, and Yong Tan. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23(1):23–41, 2012.
79. Alistair Sutcliffe, Robin I M Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British journal of psychology*, 103(2):149–168, 2012.

80. Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.
81. Schelling Thomas. Micromotives and macrobehavior, 1978.
82. Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.
83. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.
84. Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.
85. Haizhou Wang and Mingzhou Song. Clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29–33, 2011.
86. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.
87. Barry Wellman. Studying personal communities. *Social structure and network analysis*, pages 61–80, 1982.
88. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.
89. Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.
90. Christo Wilson, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web (TWEB)*, 6(4):17, 2012.
91. Shaozhi Ye and S Felix Wu. Measuring message propagation and social influence on twitter.com. In *Social informatics*, pages 216–231. Springer, 2010.
92. W-X Zhou, Dider Sornette, Russell A Hill, and Robin I M Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 272(1561):439–444, 2005.

A

Calculation of the a_k Values

In order to set a_k constants properly, we leverage on the real growth trend of Facebook over time. Hence, we approximate the Facebook network's evolution reported in [89] with the piecewise function $g(t)$ defined as:

$$g(t) = \begin{cases} 8,876,376 - 720,099 \cdot t \text{ if } t < 10 \\ 3,348,056 - 167,267 \cdot t \text{ if } 10 \leq t < 18 \\ 580,070 - 13,490 \cdot t \quad \text{if } t \geq 18 \end{cases} \tag{A.1}$$

where t is the time in months before the time of the crawl. The first elbow point of the function is placed 18 months before the time of the crawl (October 2006), when Facebook opened to everyone. Before that time, the membership was restricted to university and high-school students only. The second elbow point is placed 10 months before the time of the crawl (February 2007), when Facebook starts to become popular and its growth trend shows a significant acceleration.

For each class of relationship C_k, let μ_k be the mean value of $g(t)$ with $t \in (w_k, w_{k-1})$ and let $\bar{d}_k$ be the point in time where $g(t)$ is equal to μ_k. Resulting values for $\bar{d}_k = g^{-1}(\mu_k)$ are: $\bar{d}_1 = 0.5$, $\bar{d}_2 = 3.5$, $\bar{d}_3 = 8.74$ and $\bar{d}_4 = 20.88$. The placement of these values over the Facebook growth function $g(t)$ is depicted in Figure A.1.

Reasonably assuming that the growth trend of the links is proportional to the growth trend of the nodes, we can consider $\bar{d}_k$ as the average duration of the relationships belonging to the class C_k. In order to force the means of estimated links duration to be equal to the means obtained by the Facebook growth function, we set the constants a_k to satisfy the following equation:

$$\frac{1}{|C_k|} \sum_{r \in C_k} \hat{d}(r) = \bar{d}_k \tag{A.2}$$

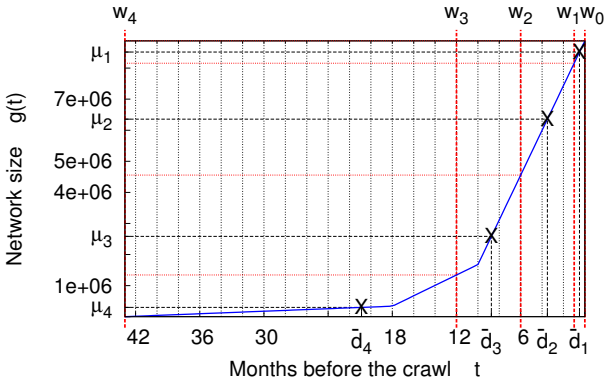

Figure A.1: The growth of Facebook over time from the time Facebook started (September 2004) to the time of the crawl (April 2008).

We obtain the following values of a_k: $a_1 = 1$, $a_2 = 3.18$, $a_3 = 3.69$ and $a_4 = 3.79$.